Università degli Studi "Roma Tre"

Scuola dottorale in "Economia e metodi quantitativi"

COVARIATE-DEPENDENT RANDOM EFFECTS IN SURVIVAL ANALYSIS

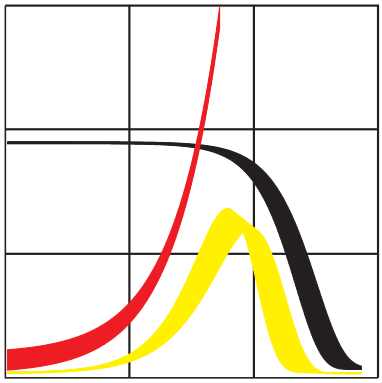
Candidato:
Francesco Cottone

Supervisore:
Prof. Francesco Lagona

XX Ciclo

SUPPORTED BY:

Max Planck Institute for Demographic Research in Rostock, Germany



Aknowledgements

This work would not have been possible without the inspiration and support provided by my supervisor Prof. Francesco Lagona of University Roma Tre. Extraordinary thanks are devoted to Dr. Jutta Gampe for the financial support she provided me on behalf of Max Planck Institute for Demographic Research in Rostock, Germany, for two years of my life.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Survival analysis studies the duration between a defined time origin and the occurrence of an event of interest. In a given population different event times reflect each subject's own susceptibility to the event itself, which is the result of a set of determinants which make the individuals different. We define such diversity as *heterogeneity*. Part of it can be described by subject-specific observable features which can be included in a survival model by a set of subject-specific observable covariates. Such models are called *fixed effects* models and account for the *observed heterogeneity*. It may be that *unobservable* features exist however, which let the individuals (or groups of them) differ with respect to the suceptibility to the event of interest. This implies the existence of an *unobserved* heterogeneity which is modelled via *random effects* which are not observed realizations of a random variable. A great variety of random effect models applied to survival analysis [1] have been developed since Vaupel, Manton, and Stallard (1979) introduced the first in this peculiar framework. Despite of the great variety of such models there is a common element which remained unchanged, that is the random effects, namely their distribution, have been always considered something which is independent of the fixed effects. In this work we introduce the assumption that an interaction between the random effects and the fixed effects may exist

[1] in survival analysis the random effect is traditionally named *frailty*.

in some cases. To allow for such interaction we introduce a new assumption about the random effect's distribution, by letting it be dependent on one or more covariates. Our assumption is general with respect to the class of survival models to which it could be applied.

As an initial step in **Chapter 2** we provide the theoretical framework of survival analysis which is necessary to get into the remainder of the work. We start from the definition of the basic concepts of survival analysis by describing one of the possible observation schemes. Such schemes are important because they originate several issues which must be carefully considered when specifying the model as we describe in Section 2.1. Namely we define there the situation of *right-censored* and *left truncated* data which is a quite common observation scheme in survival analysis. In the remainder of the chapter we present two very popular parametric survival models (Section 2.2) which we will use in the remainder of the work. Afterwards, in Section 2.3 we explain how to introduce a set of subject-specific covariates in a survival model to account for observed heterogeneity by introducing the *proportional hazards* model (PH) which is our reference model.

Chapter 3 deals with unobserved heterogeneity. At first we describe in Section 3.1 how it originates by introducing the mixture models and we describe also how the presence of unobserved heterogeneity affects the estimation of survival models' quantities of interest. In Sections 3.1 and 3.3 the original frailty model is described and a generalized survival random effects model are described which are further steps we need to specify our model.

In **Chapter 4** we describe the theory of the Expectation Maximization (EM) and the Monte Carlo EM algorithm (MCEM). These are the estimation procedures we chose to use in this work because both of them consent us to iteratively achieve maximum likelihood estimates of models with *incomplete* data sets. The presence of unobserved components in survival models originates an incomplete data set to deal with. The MCEM algorithm is basically an EM algorithm where a Monte Carlo integration is required to approximate an expected value. In Section 4.2 we show how we may sample the values we

need for such approximation when it is not possible to draw them from the distribution of interest directly. Section 4.3 describes a sampling procedure which under certain conditions allows us to draw samples from a multivariate distribution in a really simple way.

Chapter 5 extends the PH model described in Section 2.3 by introducing missing values in the covariates and explains how to estimate it via the EM algorithm, as suggested by Herring and Ibrahim (2001). This is a necessary step for the specification of the estimation procedure of our model in the remainder of the work. In this chapter we explain at first the possible different kinds of missingness generating processes (MGP). Then, in Section 5.1 we specify the PH model with missing values (PHMV) given the missing values have been generated along a specific MGP, allowing for right-censoring and left-truncation as well. In Sections 5.2 and 5.3 we describe in detail how to estimate the the PHMV model via EM and MCEM algorithms, when the covarites with missing values are discrete and continuous respectively.

We explain our original assumption on the random effects' distribution in **Chapter 6**. In the same chapter we introduce this assumption in proportional hazards model with random effects and we explain how to estimate such model. After describing why such assumption could be reasonable we briefly recall the traditional random effect PH model, we specify our assumption formally in Section 6.2. Afterwards we introduce such assumption in a proportional hazards model with random effects, allowing for both right-censoring and left-truncation as well. This is in Section 6.3. In Sections 6.3.1 and 6.3.2 we describe how to estimate such model via the EM and MCEM algorithm assuming the random effect to be discrete and continuous respectively.

Chapter 7presents the results of two main sets of simulations both implemented by assuming our assumption was true. The main purpose of them was to test the performances of the estimation procedures of the model described in the previous chapter. The two sets of simulations differs by the nature of the random effects which are assumed to be binary in the first

set, continuous in the second. In order to estimate the model accounting for unobserved heterogeneity we exploited the EM and MCEM algorithm respectively. In both simulations sets we ran 100 different independent simulations in order to achieve stable results. Moreover, in each set we compared the EM and MCEM performances to those of two other models. The first one is miss-specified, that is we specified it as if there was not unobserved heterogeneity, i.e. we did not include a random effect in it. The second model is over-specified, as we assumed to know the exact value of random effects. Both these last two models were estimated by maximizing the respective likelihoods. The results show good performances of both EM and MCEM algorithm with respect to the over-specified model while the miss-specified model shows how the estimates are distorted if we do not account for unobserved heterogeneity.

In addition to theoretical statistical modelling this work relies heavily on computer programmed algorithms necessary for simulations. Each model simulated in Chapter 7 has been implemented in software R and is largely made up of original and self-written code.

Chapter 2

Theoretical framework

Survival analysis is a branch of statistics which deals with duration data, aiming at analyzing the timing of events. Although statistical methods for duration data had been developed in engineering and biomedical sciences mainly, a wide literature spread in economics as well. For example survival analysis methods are largely applied in studies about unemployment[1], lifecycle of firms[2] and finance[3]. A synthetic exposition of survival analysis' main concepts and models is in Kiefer (1988), while an extended review of methods for the study of duration data in economics can be found in Lancaster (1992). Cox and Oakes (1984), Klein and Moeschberger (1997), Hougaard (2001) are valid references to deepen theoretical and methodological issues of survival analysis.

The main variable in survival analysis is *time*, as the focus is on the duration up to a specific event. Such event could be death, failure in mechanical features, criminal acts, divorce, finding a job, etc.. In order to define a duration we need to fix a starting point from which the chronometer starts till the event's occurrence. Such starting point is named *time origin* , defined as the individual time point from which a subject begins to be exposed to the

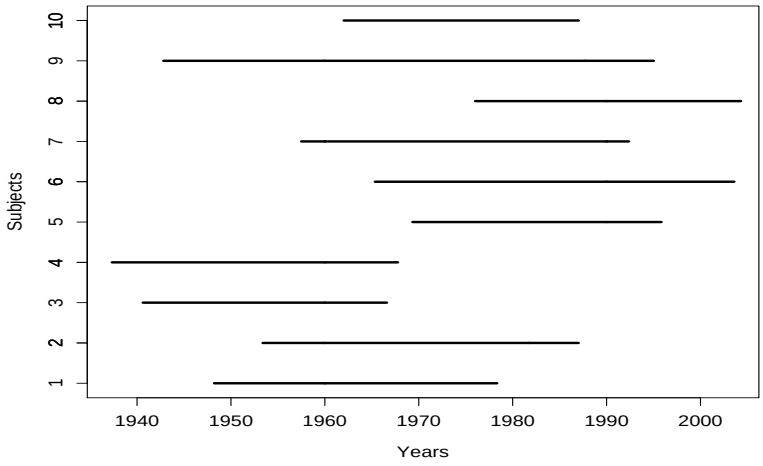
[1]Lancaster (1979); Royston (1983); Gamerman and West (1987); Belzil (2001); Cockx and Ridder (2001).

[2]Mata and Portugal (1994); Agarwal and Audretsch (2001); Agarwal and Gort (2002).

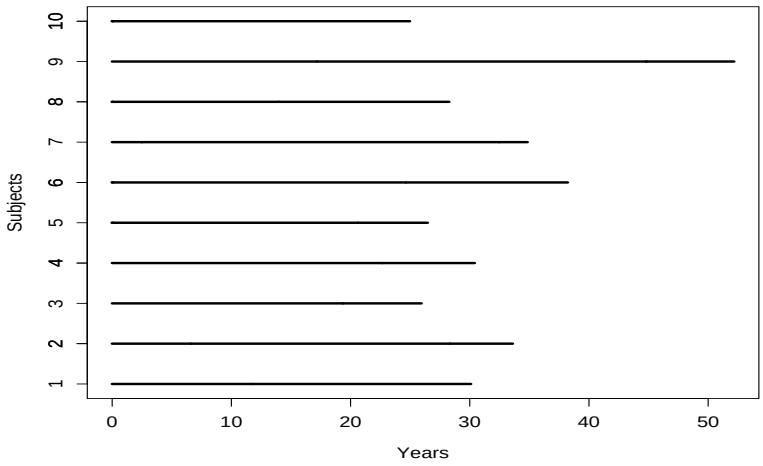
[3]Bennett et al. (2001); Guiso and Jappelli (2002); Li and Xu (2002).

risk of experiencing the event. The fixing of time origin needs some care. Sometimes it is obvious, for example if we want to study the mortality of a cohort of people. In this case the event of interest is death and it is natural to fix time origin at birth. Sometimes the definition of time origin is more difficult. Let us assume we want to study the time up to first birth. We could consider the woman at risk of giving birth from a biological point of view, i.e. since she had her first period. However, we also could fix time origin along with a behavioural assumption, e.g. at that point in time from which a woman decides to have children. In general, the choice of time origin must be carefully determined by researchers along with the study-specific theoretical hypothesis. Whatever the choice's criterion is, time origin always equals 0 for each observed subject. We name *survival* time the spell since time origin up to the event, and *failure* time the point in time at which the event occurs. When survival data are collected in practice they are represented over *calendar* time, as shown in Figure 2.1(a). Such data must be expressed as durations since a time origin to be modeled, as shown in Figure 2.1(b). Here we see that all time origins equal 0, although they are located in different years when represented over calendar time.

An implicit assumption made in Figure 2.1 is that all failure times were *fully* observable. In practice this may not be possible for some subjects, that is failure times may be *partially* observable. The lapse of time within which a subject is kept under observation is named *follow-up*. The starting point of follow-up is the individual time point since which a subject is entered the study, named *entry time*. The ending point of follow-up is named *exit time* and is the time point at which the subject exits the study. Usually failure times and follow-up times do not coincide so that a variety of possible observation schemes may occur in the same study. An example is shown in Figure 2.2 where we assume a study is performed since 1960 up to 1990 on ten subjects. Dotted lines represent *not observed* time at risk, while solid lines are follow up times. We first focus on those people who exit the study *before* they experience the event of interest, e.g subjects number eight and nine. Failure



(a) Calendar time.

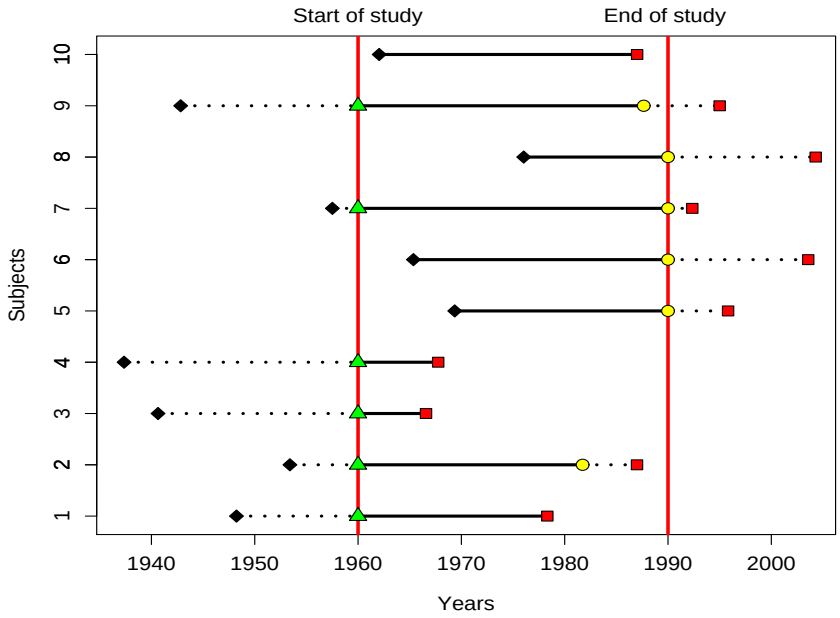


(b) Time up to event.

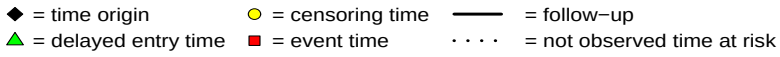
Figure 2.1: Survival times represented over calendar time (a) and as durations since time origin (b).

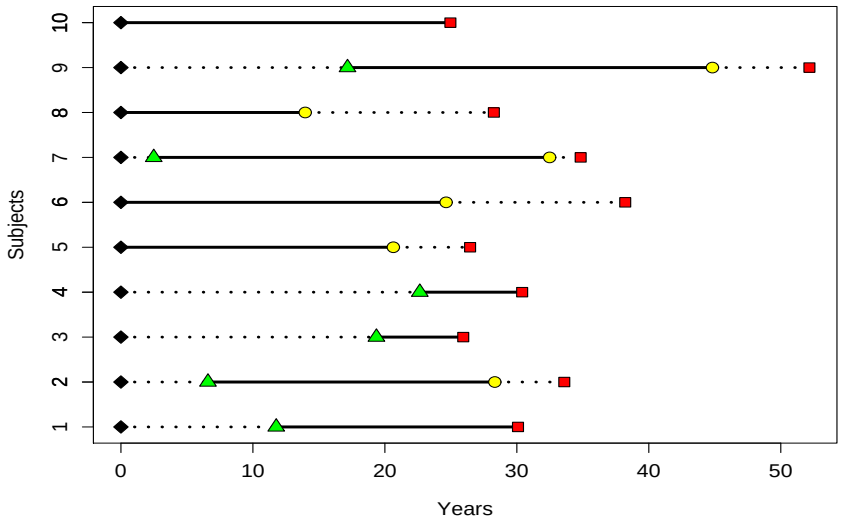
times of such subjects are named *right censored*, their observation ending at a time point named *censoring time*, which is before the event occurrence. A typical reason for right censoring to be present is a prespecified ending time of study, such that some subjects may experiment the event of interest after it, as subjects number eight, five, six and seven in Figure 2.2. In this case there is a *fixed* censoring. Another cause of right censoring is that some subjects drop the study for some reason, as subjects number two and nine in the same

figure. This is named *random* censoring. A second feature which is frequent in survival studies is *left truncation*, which occurs when subjects enter the study when they are at risk already, that is they have *delayed* entry times. For example, in Figure 2.2 failure times of those subjects who are already at risk in 1960 are left truncated (e.g. number one, three, nine). Such subjects are entered the study when they are at risk already, so that duration since time origin up to entry time is not observed. Different combinations of left truncation and right censoring are possible. For example, subject number three in Figure 2.2 is left truncated but not censored. Subjects two and nine are both left truncated and right censored. Subject eight is right-censored only, while number ten is neither truncated nor censored. The impact of left truncation and right censoring in modelling survival data will be described in Section 2.1. Here we mention that other types of truncation and censoring (and their combination) are possible, referring to Klein and Moeschberger (1997, pp. 55-65) for their description. In the remainder of this work we deal with right censoring and left truncation only and we will use the word "time" to denote a duration since time origin, as in Figure 2.2(b).



(a) Calendar time.





(b) Time up to event.

Figure 2.2: Survival times represented over calendar time (a) and as durations (b), according to an observation scheme with left-truncation and right censoring.

2.1 Basic quantities

Let $x \in [0, +\infty)$ be the possible realization of a continuos non negative random variable X. Let also $F(x) = P(X \leq x)$ and $f(x) = \frac{dF(x)}{dx}$ be the cumulative distribution function and density of X respectively. We define the *survival function* as

$$S(x) = P(X > x) = 1 - F(x) = \int_x^{+\infty} f(u)du,$$

such that

1. $S(0) = 1$, $\lim_{x \to +\infty} S(x) = 0$,

2. $S(x) > S(x')$, $\forall x < x'$, $x, x' \in [0, +\infty)$,

3. $\lim_{x \to x_0} S(x) = S(x_0)$, $\forall x_0 \in (0, +\infty)$,

4. $\lim_{x \to 0^+} S(x) = 1$,

5. $\frac{d}{dx} S(x) = \frac{d}{dx}[1 - F(x)] = -f(x)$, $\forall x$.

The survival function describes the probability that the event of interest will occur *after* time x. this implies that such occurrence is not possible at $x = 0$, which implies $S(0) = 1$. Let us define the *hazard function* as well,

$$h(x) = \lim_{\Delta x \to 0} \frac{P(x < X \leq x + \Delta x | \, X > x)}{\Delta x}, \quad \Delta x > 0, \tag{2.1}$$

which expresses the risk that event occurs in the next instant *after* x, conditional on survival to that time. Although it is always $h(x) \geq 0$, the hazard function can be differently shaped().

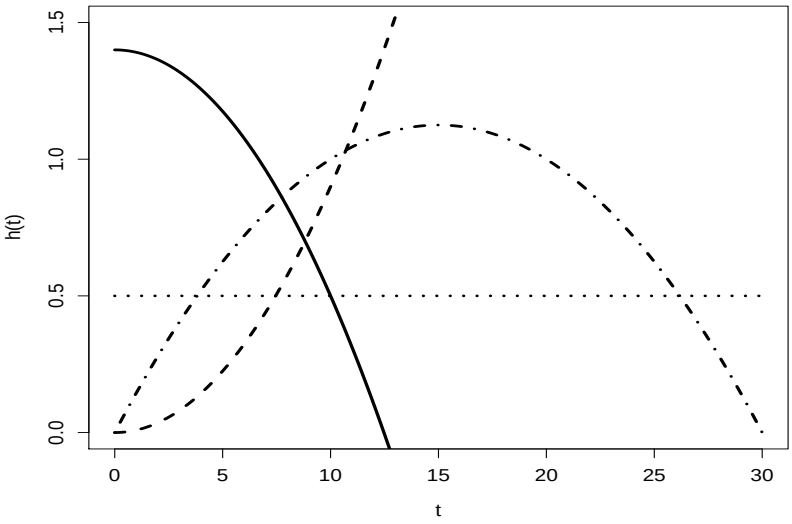


Figure 2.3: Shapes of hazard functions. Constant hazard ($\cdots\cdots$), increasing hazard ($- - -$), decreasing hazard (———), humpshaped hazard ($-\cdot-\cdot-$).

We may write (2.1) as

$$\begin{aligned} h(x) &= \lim_{\Delta x \to 0} \frac{P(x < X \leq x + \Delta x | \, X > x)}{\Delta x} \\ &= \lim_{\Delta x \to 0} \frac{P(x < X \leq x + \Delta x \cap X > x)/P(X > x)}{\Delta x} \\ &= \lim_{\Delta t \to 0} \frac{P(x < X \leq x + \Delta x)/P(X > x)}{\Delta x} \\ &= \frac{1}{P(X > x)} \lim_{\Delta x \to 0} \frac{P(x < X \leq x + \Delta x)}{\Delta x} \\ &= \frac{f(x)}{S(x)} = -\frac{d}{dx} \ln S(x). \end{aligned} \tag{2.2}$$

Hence, we define the *cumulative hazard* function

$$H(x) = \int_0^x h(u)du = -\ln S(x) \tag{2.3}$$

as that function which describes how the risk of event occurrence cumulates from time origin up to x, an instant before the event occurs.

Let $t \in \mathbb{R}^+$ denote the failure time, that is the time at which the event of interest occurs, and let us assume we observed n subjects randomly sampled from a population and the data we observed are partially right-censored and left-truncated. Each observed individual is assigned a failure time t_i, $i = 1, ..., n$, where $t_1, t_2, ..., t_n$ were independently generated by a random variable $T > 0$ along with the hazard function $h(t|\boldsymbol{\theta}) = f(t|\boldsymbol{\theta})S(t|\,\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector of parameters. To allow for right-censoring we define a positive continue random variable C with density $f_c(c|\boldsymbol{\psi})$ and cdf $F_c(c|\boldsymbol{\psi})$, where $\boldsymbol{\psi}$ is a vector of parameters, from which n i.i.d censoring times c_i are randomly drawn, $i = 1, 2, ..., n$. Furthermore, we define a positive random variable W from which n i.i.d entry times w_i are randomly drawn in order to allow for left-truncation. Each subject is therefore described by the triple (t_i, c_i, w_i). Let us recall that what we actually observe about each subject on study are an entry time w_i and an exit time which could be either an event time t_i or a censoring time c_i. We may define therefore a new positive random variable $Y = \min(T, C)$ from which n exit times $y_i = \min(t_i, c_i)$ are independently drawn, and an event indicator δ_i which equals 1 if the event occurs, 0 otherwise. Each subject is then described by the triple (y_i, δ_i, w_i). Let us assume T, C, W are mutually independent. The probability of event time t_i

conditional on surviving at time w_i may then be written as

$$\begin{aligned} P(Y \leq y_i \cap \delta_i = 1 | W > w_i) &= P(T \leq y_i \cap C > y_i \cap W > w_i)/P(W > w_i) \\ &= P(T \leq y_i \cap C > y_i)/P(W > w_i) \\ &= P(T \leq y_i)P(C > y_i)/P(W > w_i) \\ &= \frac{F_t(y_i|\boldsymbol{\theta})[1 - F_c(y_i|\boldsymbol{\psi})]}{S(w_i|\boldsymbol{\theta}_i)} \\ &= \int_0^{y_i} \frac{f(t|\boldsymbol{\theta})[1 - F_c(t|\boldsymbol{\psi})]}{S(w_i|\boldsymbol{\theta}_i)} dt. \end{aligned} \tag{2.4}$$

Similarly, the probability of censoring time c_i *conditional* on surviving at time w_i may be expressed by

$$\begin{aligned} P(Y \leq y_i \cap \delta_i = 0 | W > w_i) &= P(C \leq y_i \cap T > y_i \cap W > w_i)/P(W > w_i) \\ &= P(C \leq y_i \cap T > y_i)/P(W > w_i). \\ &= P(C \leq y_i)P(T > y_i)/P(W > w_i). \\ &= \frac{F_c(y_i|\boldsymbol{\psi})S(y_i|\boldsymbol{\theta})}{S(w_i|\boldsymbol{\theta}_i)} \\ &= \int_0^{y_i} \frac{f_c(c|\boldsymbol{\psi})S(c|\boldsymbol{\theta})}{S(w_i|\boldsymbol{\theta}_i)} dc \end{aligned} \tag{2.5}$$

Exploiting (2.4) and (2.5) we may express the density of $y_i = \min(t_i, c_i)$ as

$$f(y_i, \delta_i | w_i, \boldsymbol{\theta}, \boldsymbol{\psi}) = \begin{cases} f(y_i|\boldsymbol{\theta})[1 - F_c(y_i|\boldsymbol{\psi})]/S(w_i|\boldsymbol{\theta}) & \text{if} \quad \delta_i = 1 \\ \\ S(y_i|\boldsymbol{\theta}) f_c(y_i|\boldsymbol{\psi})/S(w_i|\boldsymbol{\theta}) & \text{if} \quad \delta_i = 0. \end{cases}$$

In order to estimate $\boldsymbol{\theta}, \boldsymbol{\psi}$ we may therefore maximize the likelihood

$$L(\boldsymbol{\theta}, \boldsymbol{\psi} | y_i, \delta_i, w_i) = \prod_i^n \left[\frac{f(y_i|\boldsymbol{\theta})(1 - F_c(y_i|\boldsymbol{\psi}))}{S(w_i|\boldsymbol{\theta})} \right]^{\delta_i} \left[\frac{S(y_i|\boldsymbol{\theta}) f_c(y_i|\boldsymbol{\psi})}{S(w_i|\boldsymbol{\theta})} \right]^{1-\delta_i}. \tag{2.6}$$

Let us assume there is *non informative* censoring, that is $\boldsymbol{\psi} \neq \boldsymbol{\theta}$. Hence, inference about $\boldsymbol{\psi}$ is superflous and (2.6) becomes

$$L(\boldsymbol{\theta} | y_i, \delta_i, w_i) = \prod_i^n \left[\frac{f(y_i|\boldsymbol{\theta})^{\delta_i} S(y_i|\boldsymbol{\theta})^{1-\delta_i}}{S(w_i|\boldsymbol{\theta})} \right]. \tag{2.7}$$

Thus, the log-likelihood corresponding to (2.7) may be written as

$$\begin{aligned} \ell(\boldsymbol{\theta} | y_i, \delta_i, w_i) &= \log \left\{ \prod_i^n \left[\frac{f(y_i|\boldsymbol{\theta})}{S(w_i|\boldsymbol{\theta})} \right]^{\delta_i} \left[\frac{S(y_i|\boldsymbol{\theta})}{S(w_i|\boldsymbol{\theta})} \right]^{1-\delta_i} \right\} \\ &= \sum_i^n \delta_i [\log f(y_i|\boldsymbol{\theta}) - \log S(w_i|\boldsymbol{\theta})] + (1 - \delta_i)[\log S(y_i|\delta_i) - \log S(w_i|\boldsymbol{\theta})] \\ &= \sum_i^n \delta_i \log f(y_i|\boldsymbol{\theta}) + (1 - \delta_i)[\log S(y_i|\delta_i)] - \log S(w_i|\boldsymbol{\theta}) \\ &= \sum_i^n \delta_i \log [h(y_i|\boldsymbol{\theta}) S(y_i|\boldsymbol{\theta})] + (1 - \delta_i)[\log S(y_i|\delta_i)] - \log S(w_i|\boldsymbol{\theta}) \\ &= \sum_i^n \delta_i \log h(y_i|\boldsymbol{\theta}) + \log S(y_i|\delta_i) - \log S(w_i|\boldsymbol{\theta}) \\ &= \sum_i^n \delta_i \log h(y_i|\boldsymbol{\theta}) - [H(y_i|\boldsymbol{\theta}) - H(w_i|\boldsymbol{\theta})], \end{aligned}$$

where

$$H(y_i|\delta_i) - H(w_i|\boldsymbol{\theta}) = \int_{w_i}^{y_i} h(t|\boldsymbol{\theta})dt.$$

Note that for those individuals who are not left-truncated the entry time concides with time origin, that is $w_i = 0$ and which implies $S(w_i) = 1$. In this case the ith log-likelihood contribution becomes $\delta_i \log h(y_i|\boldsymbol{\theta}) - H(y_i|\boldsymbol{\theta})$.

2.2 Parametric models for survival data

Different assumptions about the distribution of failure time t lead to different models for survival data. The first model we describe is the *exponential* distribution (Figure 2.4). The density function is $\lambda \exp(-\lambda t)$, $\lambda > 0$, with expected value $1/\lambda$ and variance $1/\lambda^2$. The survival function is given by

$$S(t) = \int_t^{+\infty} \lambda e^{-\lambda t} = e^{-\lambda t}.$$

The exponential random variable is the only one which is memoryless, because of the so called lack of memory property

$$\begin{aligned} P(T > t + s|T > t) &= \frac{P(T > t + s \cap T > t)}{P(T > t)} \\ &= \frac{P(T > t + s)}{P > t} = \frac{e^{-\lambda(t+s)}}{e^{-\lambda(t)}} \\ &= e^{-\lambda t - \lambda s + \lambda t} \\ &= e^{-\lambda s} = P(T > s), \qquad \forall t, s,\ s > 0. \end{aligned}$$

This property implies an important consequence about $E(T - t|T > t)$, i.e. the expected value of duration up to the event conditional on surviving at time t. Such quantity is named *mean residual life* if the failure times'

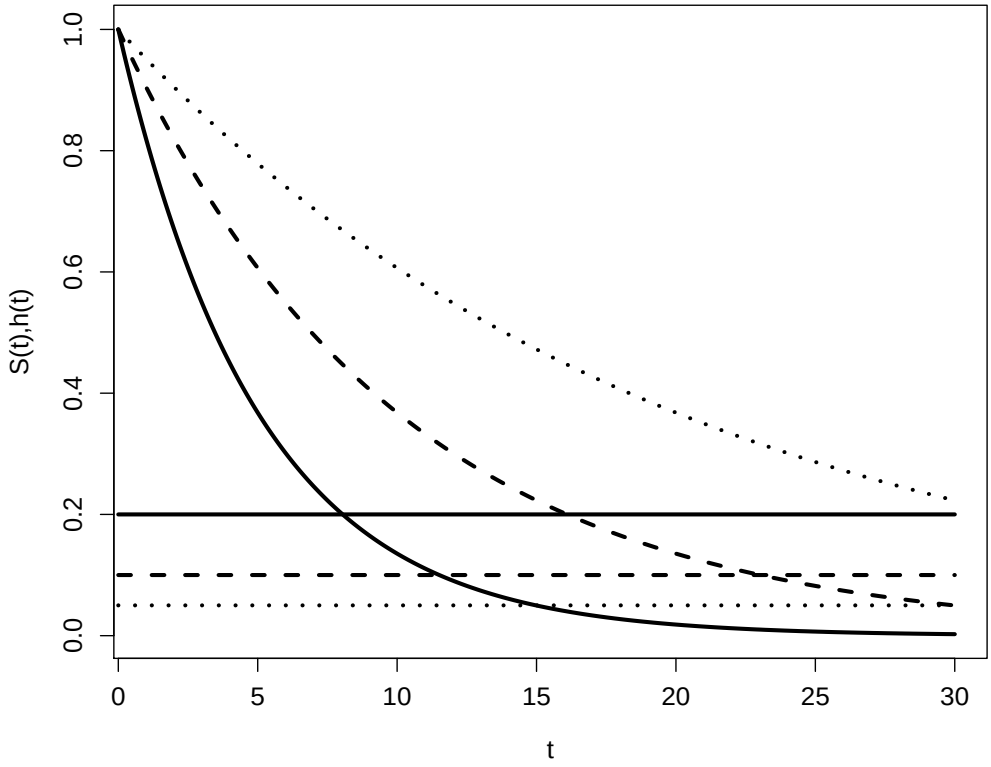


Figure 2.4: Exponential survival and hazard functions for $\lambda = 0.2$ (———), $\lambda = 0.1$ (– – –), $\lambda = 0.05$(· · · · · ·).

distribution is exponential it is

$$\int_t^{+\infty} \frac{(u-t)\lambda \exp(-\lambda u)}{\exp(-\lambda t)} du = \frac{1}{\lambda}$$

which is constant over time, that is the residual time up to the event does not depend on past history. Such "no aging" property is also reflected in the constant hazard rate

$$h(t) = -\frac{d}{dt} \ln S(t) = -\frac{-\lambda \exp(-\lambda t)}{\exp(-\lambda t)} = \lambda.$$

Although the exponential distribution is widely used for its simplicity and tractability, its constant hazard rate limits its applicability to many realistic applications.

The next distribution was originally proposed in specific studies (Rosin and Rammler, 1933; Weibull, 1939), but it was Weibull (1951) to show how it can be exploited to model a wide range of problems in different disciplines. The *Weibull* distribution is the landmark for a variety of models (Prabhakar Murthy et al., 2003). Although different formulations are available, in the remainder of this work we will refer to the two-parameters Weibull distribution, with density function

$$\alpha\lambda t^{\alpha-1}e^{-\lambda t^{\alpha}} \tag{2.8}$$

where $\alpha, \lambda \in \mathbb{R}^+$ are the shape and scale parameters respectively[4]. The rth moment of the Weibull distribution is $[\Gamma(1 + r/\alpha)]\lambda^{-r/\alpha}$ where $\Gamma(\cdot)$ is the gamma function. The mean and variance are $[\Gamma(1 + 1/\alpha)]\lambda^{-1/\alpha}$ and $\{\Gamma(1+2/\alpha)-[\Gamma(1+1/\alpha)]^2\}\lambda^{-1/\alpha}$ respectively. The Weibull survival function is given by

$$e^{-\lambda t^{\alpha}} \tag{2.9}$$

and the hazard function is

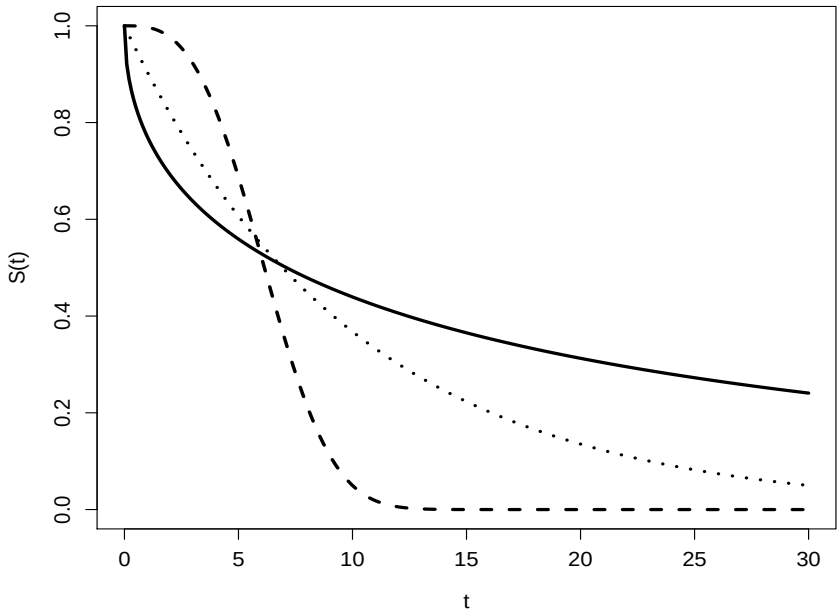
$$\alpha\lambda t^{\alpha-1}. \tag{2.10}$$

The flexibility of Weibull distribution is evident in Figure 2.5, where different possible survival and hazard functions are plotted according to different values of α and λ in (2.9) and (2.10) respectively. Note that the exponential distribution is a particular case of the Weibull distribution when $\alpha = 1$.

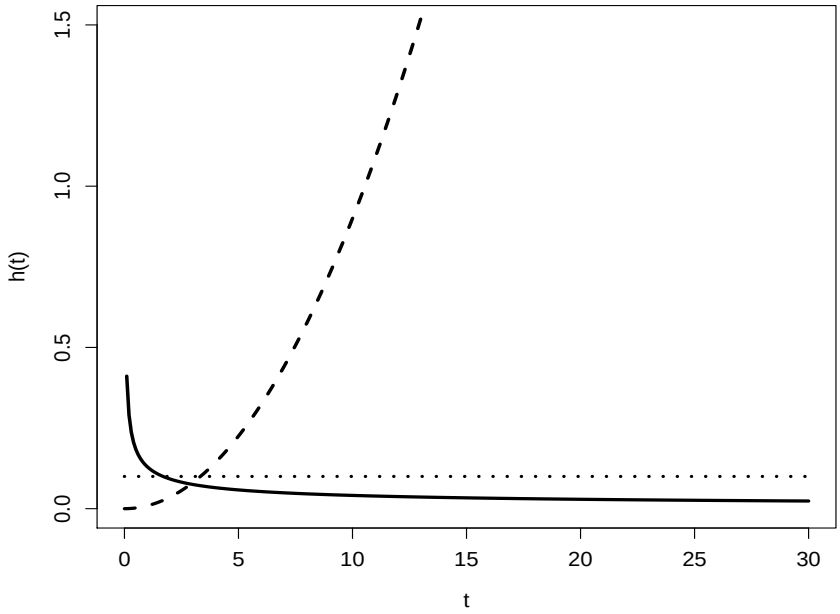
[4]The general form of Weibull density function is

$$\frac{\alpha}{\gamma}\left(\frac{t-\tau}{\gamma}\right)^{\alpha-1}\exp\left[-\left(\frac{t-\tau}{\gamma}\right)^{\alpha}\right], \qquad \alpha, \gamma, \tau \in \mathbb{R},\ \alpha > 0,\ \gamma > 0,\ t \geq \tau,$$

where α, γ, τ, are the shape, scale and location parameters respectively. The density function (2.8) is obtained setting $\tau = 0$ and $\gamma^{-\alpha} = \lambda$.



(a) Survival functions.



(b) Hazard functions.

Figure 2.5: Weibull survival and hazard functions for $\alpha = 0.5, \lambda = 0.26$ (———), $\alpha = 1, \lambda = 0.1$ (– – –), $\alpha = 3, \lambda = 0.003$(· · · · · ·).

2.3 Regression for survival data

So far we assumed that all the variability of failure times derived from the realizations of a random variable T, i.e. the population from which the *n-sample* was drawn was *homogeneus* with respect to the hazard $h(t)$. Formally, we guessed that all the observed subjects share the *same* hazard function, but this assumption seems to be reasonably too restrictive. The individuals could differ one from each other according to some own specific features which may have an impact on the individual survival chances, like sex, socio-economic conditions etc.. Such additional subject-specific information may be summarized via a vector of *observable* covariates $\boldsymbol{x} = [x_1, x_1, ..., x_k]'$, whose value is subject-specific as well. The individual covariates' profiles account for the *observed* heterogeneity, and can be either *time constant* or *time varying*. In the remainder of this work we deal with time constant covariates only.

Different assumptions can be made on the way the covariates impact on the individual duration, leading to different models (Klein and Moeschberger, 1997). A common feature of such models is that they all deal with a *subject-specific conditional* hazard rate $h(t|\boldsymbol{x})$, which is the result of a link between the individual covariates' profile and a *baseline* hazard rate $h_0(\cdot)$. The baseline hazard accounts for a shared component of risk which is the same for all the members of a population.

Let $\boldsymbol{x}_i$ be the individual specific k-vector of covariates of the ith subject, $i = 1, 2, ..., n$. The basic assumption of proportional hazards models (PH) is that $\boldsymbol{x}_i$ impacts *multiplicatively* on the ith hazard rate $h(t|\boldsymbol{x}_i)$ via a *non negative* function of the covariates $g(\boldsymbol{x}_i)$, that is

$$h(t|\boldsymbol{x}_i) = h_0(t)g(\boldsymbol{x}_i), \qquad g(\cdot) > 0,\ g(\boldsymbol{x}_i = \mathbf{0}) = 1, \tag{2.11}$$

where $h_0(t)$ is the baseline hazard which represents the risk when $\boldsymbol{x}_i = 0$.

Thus, recalling (2.3) and (2.2), the survival and the density functions are

$$S(t|\boldsymbol{x}_i) = \exp[-H_0(t)g(\boldsymbol{x}_i)] \tag{2.12}$$

and

$$f(t|\boldsymbol{x}_i) = h_0(t)g(\boldsymbol{x}_i)\exp[-H_0(t)g(\boldsymbol{x}_i)]$$

respectively.

Assuming that the value of $\boldsymbol{x}_i$ is fixed at time origin $\forall i$, a key feature of PH models is that the hazard rates ot two subjects are proportional and constant over time, that is

$$\frac{h(t|\boldsymbol{x}_i)}{h(t|\boldsymbol{x}_j)} = \frac{h_0(t)g(\boldsymbol{x}_i)}{h_0(t)g(\boldsymbol{x}_j)} = \frac{g(\boldsymbol{x}_i)}{g(\boldsymbol{x}_j)}. \tag{2.13}$$

where $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$,$\boldsymbol{x}_i \neq \boldsymbol{x}_j$ are the covariates profiles of two subjects. The quantity (2.13) is the relative risk of the subject with covariate profile $\boldsymbol{x}_i$ with respect to the individual with covariate profile $\boldsymbol{x}_j$. The concept of multiplicative impact of the covariates on the baseline hazard was first introduced by Cox (1972). In this seminal paper the link function is

$$g(\boldsymbol{x}_\cdot) = \exp(\boldsymbol{\beta}'\boldsymbol{x}_\cdot) \tag{2.14}$$

where $\boldsymbol{\beta}$ is a vector of parameters. The hazard function is given by

$$h(t|\boldsymbol{x}_\cdot) = h_0(t)\exp(\boldsymbol{\beta}'\boldsymbol{x}_\cdot).$$

Under the assumption of time constant covariates, the relative risk (2.13) becomes

$$\frac{h(t|\boldsymbol{x}_i)}{h(t|\boldsymbol{x}_j)} = \frac{h_0(t)\exp(\boldsymbol{\beta}'\boldsymbol{x}_i)}{h_0(t)\exp(\boldsymbol{\beta}'\boldsymbol{x}_j)} = \exp[\boldsymbol{\beta}'(\boldsymbol{x}_i - \boldsymbol{x}_j)].$$

Thus, if we apply (2.14) to (2.11) and (2.12) the survival and the density

functions are

$$S(t|\boldsymbol{x}_i) = \exp[-H_0(t)\exp(\boldsymbol{\beta}'\boldsymbol{x}_i)]$$

and

$$f(t|\boldsymbol{x}_\cdot) = h_0(t)\exp(\boldsymbol{\beta}'\boldsymbol{x}_\cdot)\exp[-H_0(t)\exp(\boldsymbol{\beta}'\boldsymbol{x}_\cdot)]$$

respectively. Let $\boldsymbol{\eta}$ be the vector of parameters of the baseline hazard function, such that $h_0(t|\boldsymbol{\eta})$. Thus, allowing for non informative right-censoring and left-truncation and recalling (2.7) the likelihood to maximize may be written as

$$L(\boldsymbol{\theta}|y_i,\delta_i,w_i) = \prod_i^n \left\{ \frac{[h_0(y_i|\boldsymbol{\eta})\exp(\boldsymbol{\beta}'\boldsymbol{x}_i)]^{\delta_i}\exp[-H_0(y_i)\exp(\boldsymbol{\beta}'\boldsymbol{x}_i)]}{\exp[-H_0(w_i)\exp(\boldsymbol{\beta}'\boldsymbol{x}_i)]} \right\}, \quad (2.15)$$

where $\boldsymbol{\theta} = [\boldsymbol{\eta}, \boldsymbol{\beta}]$.

Chapter 3

Unobserved heterogeneity

So far we described survival models in which the risk of failure is due to a combination of two elements. The first one is given by the baseline, a part of risk which is ruled by the same distribution of a positive random variable T. The other component of risk is subject-specific and impacts on the hazard rate via the individual covariates' profiles, which capture the observed heterogeneity. There could be individual (or subgroup) specific determinants however, which add an extra-heterogeneity among the individuals which is not caught by the observable covariates. Such *unobserved heterogeneity* has to be included in the model specification, in order to avoid to overestimate the survival function and to underestimate the hazard function (see Omori and Johnson, 1993). In this chapter we describe how the unobserved heterogeneity can be modeled via *mixture models*.

3.1 Mixture models

Mixture models are useful to describe situations in which a population is parted in s different subgroups, such that the observations follow a distribution which is a mixture of s subgroup-specific distributions named *compo-*

nents,

$$f(x) = \sum_{j=1}^{s} \pi_j \cdot f_j(x),$$

where $\pi_j > 0$ is the proportion of the jth subgroup with distribution $f_j(x)$, $\sum_{j=1}^{s} \pi_j = 1$. The vector $\boldsymbol{\pi} = [\pi_1, ..., \pi_s]'$ is named *mixing distribution* as it provides the probabilities of membership to the jth subgroup of a subject belonging to the observed population. In a survival analysis framework the presence of subpopulations with different distributions of failure times leads to important consequences on the estimation of survival and hazard functions. Let us assume that at time $t = 0$ we have a population of n subjects described by a survival function $S(t)$ and an hazard function $h(t)$. Let Z be a binary random variable which can be equal to 1 or 2, and n_1 and n_2, $n = n_1 + n_2$, the number of those subjects with $z = 1$ and $z = 2$ respectively. In other words, there are $s = 2$ subgroups in the population according to the values that Z can have. Let us also assume that such subgroups have a different susceptibility to failure, namely

$$h_1(t|Z = 1) < h_2(t|Z = 2) \tag{3.1}$$

which are *conditional* hazards, as they rule the survival times of those individuals *within* subgroup 1 or subgroup 2 respectively. We denote $S_j = S(t|Z = j)$ and $f_j = f(t|Z = j)$ as the conditional survival and density function of the jth subgroup, $j = 1, 2$. The *marginal* survival function of the whole population is given by

$$\begin{aligned} S(t) &= P(T > t) = P[(T > t) \cap (Z = 1)] + P[(T > t) \cap (Z = 2)] \\ &= P(T > t|Z = 1) \cdot P(Z = 1) + P(T > t|Z = 2)P(Z = 2) \\ &= S_1(t) \cdot \pi_1(0) + S_2(t) \cdot \pi_2(0) \end{aligned}$$

where $\pi_1(0)$ and $\pi_2(0) = 1 - \pi_1(0)$ are the proportions *at time* 0 of subjects with $Z = 1$ and $Z = 2$ respectively. Assuming that all subjects are at risk since $t = 0$, the proportions of the two subgroups will change over time according to a *selection* effect. Namely, the individuals with $Z = 2$ will fail earlier than those belonging to subgroup 1, because of assumption (3.1). The subgroups' proportions at time $t > 0$ will be

$$\begin{aligned}
\pi_1(t) &= P(Z = 1|T > t) = \frac{P(Z = 1 \cap T > t)}{P(T > t)} \\
&= \frac{P(T > t|Z = 1)P(Z = 1)}{P(T > t)} \\
&= \frac{S_1(t)\pi_1(0)}{S(t)} \\
\pi_2(t) &= \frac{S_2(t)\pi_2(0)}{S(t)}, \qquad (3.2)
\end{aligned}$$

and the marginal density at time t is

$$\begin{aligned}
f(t) &= -\frac{d}{dt}S(t) = -\frac{d}{dt}[S_1(t) \cdot \pi_1(0) + S_2(t) \cdot \pi_2(0)] \\
&= S_1'(t) \cdot \pi_1(0) + S_2'(t) \cdot \pi_2(0) \\
&= f_1(t)\pi_1(0) + f_2(t)\pi_2(0).
\end{aligned}$$

Hence, the marginal hazard at time t is

$$\begin{aligned}
h(t) &= \frac{f(t)}{S(t)} = \frac{f_1(t)\pi_1(0)}{S(t)} + \frac{f_2(t)\pi_2(0)}{S(t)} \\
&= \frac{f_1(t)\pi_1(t)}{S_1(t)} + \frac{f_2(t)\pi_2(t)}{S_2(t)} \\
&= h_1(t)\pi_1(t) + h_2(t)\pi_2(t) \qquad (3.3)
\end{aligned}$$

because, exploiting (3.2),

$$\frac{\pi_j(0)}{S(t)} = \frac{\pi_j(t)}{S_j(t)}, \qquad j = 1, 2.$$

The hazard function of the population is therefore the weighted average of the conditional hazards of the subgroups, as shown in (3.3). The subjects belonging to subgroup 2 will tend to fail earlier than those of group 1 however, as we assumed $h_1(t|Z = 1) < h_2(t|Z = 2)$. This implies that $\pi_1(t)$ will increase over time while $\pi_2(t)$ will decrease, so that the marginal hazard (3.3) will be increasingly affected by the proportion of stronger individuals, namely those with $Z = 1$. Hence, if we do not account for the presence of the two existing subgroups, $h(t)$ would be the more underestimated the more $t \to +\infty$. The solution to this problem is trivial if we know the values of Z for each observed subjects, i.e. if we would know the initial proportions $\pi_1(0)$ and $\pi_2(0)$ exactly. The matter is how to account for the presence of subgroups when it is not possible to observe their proportions.

In general, assuming the existence of $s > 2$ subgroups we may write the marginal quantities as

$$S(t) = \sum_{j=1}^{s} \pi_j \cdot S_j(t)$$

$$\pi(t) = \sum_{j=1}^{s} \frac{\pi_j(0) \cdot S_j(t)}{S(t)}$$

$$f(t) = \sum_{j=1}^{s} \pi_j(t) \cdot f_j(t)$$

$$h(t) = \sum_{j=1}^{s} \pi_j(t) \cdot h_j(t)$$

3.1.1 Random effects

We can extend the finite mixture model allowing for Z to be a continuous random variable (RV) with density $g(z)$. The RV Z is a *random* effect, a statistical methodology widely used to model unobserved subject-specific features which make individuals different. Random effects allow us to include the extra-heterogeneity added by these unobervable features in the statistical models in order to make valid inferences. Referring to the mixture model previously described, $g(z)$ is a *continuous* mixing distribution. In a population of n subjects each individual is assigned a unique and specific value z_i, $i = 1, 2, ..., n$, which is the random outcome of Z. The zs' are unknown but their value is not required to specify the marginal quantities, which are those relevant for us. This is possible by *integrating out* the conditional quantities, which implies we need to specify a model for $g(z)$. In such a framework, the marginal survival function is given by

$$S(t) = \int_z S(t|z)g(z)dz \tag{3.4}$$

which is integrated out over the possible values of Z. The marginal hazard is

$$h(t) = \int_z h(t|z)g(z|T > t)dz \tag{3.5}$$

where the conditional density $g(z|T > t)$ is

$$\begin{aligned} g(z|T > t) &= \frac{g(z \cap T > t)}{P(T > t)} \\ &= \frac{P(T > t|z)g(z)}{P(T > t)} \\ &= \frac{S(t|z)g(z)}{S(t)}. \end{aligned}$$

3.2 Frailty model

In survival analysis the unobserved heterogeneity is modeled by a random effect called *frailty*. The name frailty recalls the fact that those individuals with higher values of z are more susceptible to failure, i.e. they are frailer. Formally, frailty is a random effect, the *unobserved* realization of a random variable which allows for including not observable determinats in the analysis. A frailty model is basically a mixture model where the random variable which clusterizes the population is assumed to be continuous. In the original frailty (*overdispersion*) model by Vaupel et al. (1979) the basic assumption is that each subject of a population is randomly assigned a specific *not observable* value $z > 0$ at birth, which affects the susceptibility to failure by acting multiplicatively on the individual hazard. Namely,the zs are *independently* drawn from a random variable $Z \sim \text{Gamma}(k, \lambda)$ with $k = \lambda$ so that $E(Z) = k/\lambda = 1$ and $\sigma_Z^2 = \kappa/\lambda^2 = 1/k$. The gamma distribution for Z was choosen because of its flexibility and the zs have to be necessarily positive as well. A frailty which equals 1 has a neutral impact on individual hazards. Let $h_0(t)$ be the baseline hazard for n subjects each assigned a $z_i > 0$ drawn from

$$Z \sim \Gamma(k, \lambda) = \frac{\lambda^k z^{k-1} e^{-\lambda z}}{\Gamma(k)}, \quad k > 0,\ \lambda > 0$$

and

$$\Gamma(k) = \int_0^\infty z^{k-1} e^{-z} dz.$$

The conditional hazard of the ith subject is therefore given by

$$h(t|z_i) = z_i \cdot h_0(t).$$

When $z_i = 1$ the ith subject is called a "standard" individual as the frailty leaves its hazard unchanged. Along with the above assumptions, (3.4) and

(3.5) become

$$\begin{aligned} S(t) &= \int_0^\infty e^{-zH(t)} \frac{\lambda^k z^{k-1} e^{-\lambda z}}{\Gamma(z)} dz \\ &= \frac{\lambda^k}{[\lambda + H_0(t)]^k} \\ &= \left[\frac{1}{1+\sigma^2 H_0(t)}\right]^{\frac{1}{\sigma^2}} \end{aligned}$$

and

$$h(t) = -\frac{d}{dt} \ln S(t) = \frac{h_0(t)}{1+\sigma^2 H_0(t)},$$

where $H_0(t)$ is the baseline hazard rate. The original frailty model has been developed along different directions. For example, Hougaard (1984, 1986) proposed several new possible distributions for the frailty discussing their properties and the consequences of their use. Petersen (1998) proposed a frailty model where the impact on hazard is given by the sum of two or more gamma frailty terms. Dynamic frailties were presented in Yue and Chan (1997) where individual random effects' values are allowed to vary over time stochastically.

3.3 PH model with random effects

In Sections 2.3 and 3.1.1 we described how to include both observed and unobserved heterogeneity in a survival model. It is straightforward to introduce a model which may account for both kinds of heterogeneity at the same time. This purpose is achieved by the proportional hazards model with random effects (PHRE). We describe this model in the formulation given in Vaida and Xu (2000), which is a generalization of the original model introduced by Clayton and Cuzick (1985).

Let a population of n subjects exist which is clusterized in P groups, each formed by n_p individuals, $p = 1, ..., P$. The ith subject of the pth cluster is

described by a k-vector of fixed covariates $\boldsymbol{x}_{p,i} = x_{p,i,1}, x_{p,i,2}, ..., x_{p,i,k}$ which is multiplied by a k-vector of parameters $\boldsymbol{\beta}$. Furthermore, each pth cluster is assigned a q-vector $\boldsymbol{f}_p$ of random effects multiplied by a group specific $n_p \times q$ matrix $\boldsymbol{\Omega}_p = [\boldsymbol{\omega}_1, ..., \boldsymbol{\omega}_{n_p}]'$. Each row $\boldsymbol{\omega}_{\cdot}$ is a q-vector of covariates which measures the impact of $\boldsymbol{f}_p$ on the hazard of the ith subject in the pth group. The random vectors $\boldsymbol{f}_1, ..., \boldsymbol{f}_p, ..., \boldsymbol{f}_P$ are assumed to be the realizations of a multivariate random variable from which they are independently drawn. In Vaida and Xu (2000) it is assumed $\boldsymbol{f}_p \sim N_q(\boldsymbol{0}, \boldsymbol{D})$, where $\boldsymbol{D}$ is a $q \times q$ covariance matrix, although such choice is not binding (see Herring et al., 2002). Note that each subject who belongs to the same group is assigned the *same* value f_p. The hazard function of the ith subject in the pth cluster is therefore given by

$$h_{p,i}(t|\boldsymbol{x}_{p,i}, \boldsymbol{\beta}, \boldsymbol{f}_p, \boldsymbol{\omega}_{p,i}) = h_0(t) \exp(\boldsymbol{\beta}'\boldsymbol{x}_{p,i} + \boldsymbol{\omega}'_{p,i}\boldsymbol{f}_p) \tag{3.6}$$

and the corresponding survival function is

$$S_{p,i}(t|\boldsymbol{x}_{p,i}, \boldsymbol{\beta}, \boldsymbol{f}_p, \boldsymbol{\omega}_{p,i}) = S_0(t) \exp(\boldsymbol{\beta}'\boldsymbol{x}_{p,i} + \boldsymbol{\omega}'_{p,i}\boldsymbol{f}_p)$$

We remark that the PHRE model allows for modelling a possible correlation of event times among subjects *within* the same group, when a clustered structure of observations is present. Note that conditional on the random effects' vector the event times in the same cluster are independent. Unconditional independence holds for event times of subjects who belong to different groups.

Now let us assume $P = n$, $p = i$, $q = 1$, that is each subject is a one-dimensional cluster which is assigned a univariate random effect f_i. If we assume also $\omega_{p,i} = 1$, $z_i = \exp(f_i)$ and $z_i \sim Gamma(k, k)$ $\forall i$, the original frailty model described in Section 3.2 can be seen as a particular case of PHRE model where $\exp(\boldsymbol{x}'_i\boldsymbol{\beta}) = 1 \, \forall i$.

Chapter 4

EM and MCEM algorithms

The expectation-maximization (EM) algorithm is an iterative computation procedure to perform maximum likelihood estimations when the data are incomplete, that is when there are missing or hidden data. Such procedure, formalized by Dempster, Laird, and Rubin (1977), is widely used to solve many inference problems which can be formulated as missing values problems, where the missing or hidden quantities may be either variables, parameters or weights (see McLachlan and Krishnan, 1996). In the next section the EM algorithm is described, while in Section 4.2.2 an EM formulation with a Monte Carlo step is presented

4.1 EM algorithm

Let $\boldsymbol{y}$ be the observed realization from a sample space $\mathscr{Y}$. We will refer to $\boldsymbol{y}$ as the *incomplete* data set. Let $\mathscr{X}$ be the sample space such that a subset $\mathscr{X}(\boldsymbol{y}) : \{\boldsymbol{x} : \boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x})\}$ exists, where the *complete* data set $\boldsymbol{x}$ is the generic realization from $\mathscr{X}$. Namely, $\mathscr{X}(\boldsymbol{y})$ is the subset of possible complete data sets which can be associated to $\boldsymbol{y}$. This implies that we only know $\boldsymbol{x}$ to lie in $\mathscr{X}(\boldsymbol{y})$. Let $p_c(\boldsymbol{x}|\boldsymbol{\theta})$ be the probability distribution function (pdf) of $\boldsymbol{x}$, such

that

$$p(\boldsymbol{y}|\boldsymbol{\theta}) = \int_{\mathscr{X}(\boldsymbol{y})} dp_c(\boldsymbol{x}|\boldsymbol{\theta}), \tag{4.1}$$

where $p(\boldsymbol{y}|\boldsymbol{\theta})$ is the pdf of $\boldsymbol{y}$. Along with (4.1) the incomplete data log-likelihood function $\ell(\boldsymbol{\theta}|\boldsymbol{y}) = \log p(\boldsymbol{y}|\boldsymbol{\theta})$ may be written as

$$\ell(\boldsymbol{\theta}|\boldsymbol{y}) = \log \int_{\mathscr{X}(\boldsymbol{y})} dp_c(\boldsymbol{x}|\boldsymbol{\theta}), \tag{4.2}$$

In order to find the maximun likelihood estimate of $\boldsymbol{\theta}$ we need a procedure which must account for the incompleteness of data (see Schafer and Graham, 2002). The basic idea of EM algorithm is to find the maximum likelihood estimates (MLE) $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ by iteratively maximizing the conditional expectation of the complete data log-likelihood function $\ell_c(\boldsymbol{\theta}|\boldsymbol{x}) = \log p_c(\boldsymbol{x}|\boldsymbol{\theta})$, given $\boldsymbol{y}$ and a current estimate $\boldsymbol{\theta}^h$,

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^h) = E_{\boldsymbol{\theta}^h}[\ell_c(\boldsymbol{\theta}|\boldsymbol{x})|\boldsymbol{y}] = \int_{\mathscr{X}(\boldsymbol{y})} \ell_c(\boldsymbol{\theta}|\boldsymbol{x}) dp(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta}^h) \tag{4.3}$$

where $p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta})$ is the *conditional* pdf of complete data $\boldsymbol{x}$ given the observed data $\boldsymbol{y}$ and $\boldsymbol{\theta}^h$. At the $(h+1)$th iteration the EM algorithm is a two-steps procedure:

1. E-step: compute $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^h)$

2. M-step: find $\boldsymbol{\theta}^{h+1} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^h)$ and set $\boldsymbol{\theta}^{h+1} = \boldsymbol{\theta}^h$.

As stated by Efron in the discussion of Dempster et al. (1977), the effectiveness of EM algorithm in maximizing $\ell(\boldsymbol{\theta})$ at each hth iteration lies in the identity

$$u(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}[u_c(\boldsymbol{\theta})|\boldsymbol{y}] \tag{4.4}$$

proved by Fisher (1925), where $u(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta}|\boldsymbol{y})$ and $u_c(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \ell_c(\boldsymbol{\theta}|\boldsymbol{x})$. The M-step finds $\boldsymbol{\theta}^{h+1}$ by solving $E_{\boldsymbol{\theta}}[u_c(\boldsymbol{\theta}|\boldsymbol{x}] = 0$, which implies that $\boldsymbol{\theta}^{h+1}$ is the solution of $u(\boldsymbol{\theta}) = 0$ as well because of (4.4).

Furthermore, the series of EM estimates $\{\boldsymbol{\theta}^0, \boldsymbol{\theta}^1, \boldsymbol{\theta}^2, ..., \}$ is such that

$$\ell(\boldsymbol{\theta}^{h+1}) - \ell(\boldsymbol{\theta}^h) \geq 0, \quad h = 0, 1, 2, ... \quad , \tag{4.5}$$

i.e the incomplete data log-likelihood is increased at each step. In order to prove this, let us consider the conditional density of the generic possible complete data set $\boldsymbol{x}$, which we may write as

$$p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta}) = p_c(\boldsymbol{x}|\boldsymbol{\theta})/p(\boldsymbol{y}|\boldsymbol{\theta}),$$

so we may write the incomplete data log-likelihood as

$$\ell(\boldsymbol{\theta}^*|\boldsymbol{y}) = \ell_c(\boldsymbol{\theta}^*|\boldsymbol{x}) - \log p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta}^*), \tag{4.6}$$

where $\boldsymbol{\theta}^*$ is a given value of $\boldsymbol{\theta}$. Hence, at the $(h+1)$th iteration of the EM algorithm we may write the conditional expectation of (4.6) as

$$\begin{aligned} \ell(\boldsymbol{\theta}^{h+1}|\boldsymbol{y}) &= E_{\boldsymbol{\theta}^h}[\ell_c(\boldsymbol{\theta}^{h+1}|\boldsymbol{x})|\boldsymbol{y}] - E_{\boldsymbol{\theta}^h}[\log p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta}^{h+1})|\boldsymbol{y}] \\ &= Q(\boldsymbol{\theta}^{h+1}|\boldsymbol{\theta}^h) - H(\boldsymbol{\theta}^{h+1}|\boldsymbol{\theta}^h), \end{aligned}$$

where $Q(\boldsymbol{\theta}^{h+1}|\boldsymbol{\theta}^h)$ is the conditional expected value of $\ell_c(\boldsymbol{\theta}|\boldsymbol{x})$ computed for $\boldsymbol{\theta} = \boldsymbol{\theta}^{h+1}$ given the starting value $\boldsymbol{\theta}^h$, and $H(\boldsymbol{\theta}^{h+1}|\boldsymbol{\theta}^h) = E_{\boldsymbol{\theta}^h}[\log p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta}^{h+1})|\boldsymbol{y}]$. Analogously, it is

$$\ell(\boldsymbol{\theta}^h|\boldsymbol{y}) = Q(\boldsymbol{\theta}^h|\boldsymbol{\theta}^h) - H(\boldsymbol{\theta}^h|\boldsymbol{\theta}^h),$$

such that we may write

$$\begin{aligned} \ell(\boldsymbol{\theta}^{h+1}) - \ell(\boldsymbol{\theta}^h) &= Q(\boldsymbol{\theta}^{h+1}|\boldsymbol{\theta}^h) - H(\boldsymbol{\theta}^{h+1}|\boldsymbol{\theta}^h) - [Q(\boldsymbol{\theta}^h|\boldsymbol{\theta}^h) - H(\boldsymbol{\theta}^h|\boldsymbol{\theta}^h)] \\ &= Q(\boldsymbol{\theta}^{h+1}|\boldsymbol{\theta}^h) - Q(\boldsymbol{\theta}^h|\boldsymbol{\theta}^h) - [H(\boldsymbol{\theta}^{h+1}|\boldsymbol{\theta}^h) - H(\boldsymbol{\theta}^h|\boldsymbol{\theta}^h)]. \end{aligned}$$

Then, $\forall h$ it is

$$\bullet \quad Q(\boldsymbol{\theta}^{h+1}|\boldsymbol{\theta}^h) - Q(\boldsymbol{\theta}^h|\boldsymbol{\theta}^h) \geq 0, \tag{4.7}$$

and

$$\bullet \quad H(\boldsymbol{\theta}^{h+1}|\boldsymbol{\theta}^h) - H(\boldsymbol{\theta}^h|\boldsymbol{\theta}^h) \leq 0, \tag{4.8}$$

where (4.7) is true because $\boldsymbol{\theta}^{h+1}$ maximizes $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^h)$. Furthermore, (4.8) is proved because, recalling Jensen's inequality[1], it is $\forall \boldsymbol{\theta}$

$$\begin{aligned} H(\boldsymbol{\theta}|\boldsymbol{\theta}^h) - H(\boldsymbol{\theta}^h|\boldsymbol{\theta}^h) &= E_{\boldsymbol{\theta}^h}[\log p(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})|\boldsymbol{y}] - E_{\boldsymbol{\theta}^h}[\log p(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}^h)|\boldsymbol{y}] \\ &= E_{\boldsymbol{\theta}^h}\left\{\log\left[\frac{p(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})}{p(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}^h)}|\boldsymbol{y}\right]\right\} \\ &\leq \log\left\{E_{\boldsymbol{\theta}^h}\left[\frac{p(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})}{p(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}^h)}|\boldsymbol{y}\right]\right\} \\ &\propto \log\int_{\mathscr{X}(y)} dp(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}) = 0, \end{aligned}$$

and (4.5) is proved because of (4.7) and (4.8). Dempster, Laird, and Rubin (1977) prove also that the sequence of EM estimates $\{\boldsymbol{\theta}^0, \boldsymbol{\theta}^1, \boldsymbol{\theta}^2, ...,\}$ converges to a maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ although it is not ensured that the corresponding log-likelihood value $\ell(\hat{\boldsymbol{\theta}})$ is a global maximum in case of multimodal likelihood.

4.2 Monte Carlo EM algorithm (MCEM)

So far we described the EM algorithm implicitely assuming that the E-step $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^h)$ was analitically tractable, although this may not be possible sometimes. In such cases it is possible to exploit the Monte Carlo integration to build up an approximated E-step, which basically consists of an average over the possible complete data sets $\boldsymbol{x}$ independently drawn from the pdf

[1]For any random variable X, it is $E[g(x)] \leq g[E(x)]$ if g(x) is a concave function.

$p(\boldsymbol{x})$. Furthermore, it may occur that $p(\boldsymbol{x})$ is not feasible for sampling. In this case, we may sample $\boldsymbol{x}$ from a generic pdf $g(\boldsymbol{x})$ which we know to be equivalent to $p(\boldsymbol{x})$ with respect to the frequencies' distribution of sampled $\boldsymbol{x}$s. Such procedure is still a Monte Carlo integration which uses Markov chains (MCMC), which is explained in 4.2.1. In 4.3 we describe how to build up a Markov chain, namely the Gibbs sampler.

4.2.1 Markov Chain Monte Carlo theory

Let $\boldsymbol{X}$ be a random vector with pdf $p(x)$ and let $\vartheta(\boldsymbol{x})$ be a generic function of $\boldsymbol{x}$. Let us assume that our focus is on computing the expectation

$$E[\vartheta(\boldsymbol{x})] = \int \vartheta(\boldsymbol{x}) dp(\boldsymbol{x}), \tag{4.9}$$

where the integration is taken over all the possible values of $\boldsymbol{x}$. Then, if (4.9) is not directly computable it can be approximated by the Monte Carlo integration

$$\hat{E}[\vartheta(\boldsymbol{x})] = \frac{1}{R} \sum_{t=1}^{R} \vartheta(\boldsymbol{x}_t) \tag{4.10}$$

where $\{\boldsymbol{x}_t\}$ is a sequence of R i.i.d. random vectors drawn from $p(\boldsymbol{x})$ and $\hat{E}[\vartheta(\boldsymbol{x})]$ is a consistent estimate of (4.9) more accurate as R increases, because of the strong law of large numbers.

It may not be possible to draw independent samples from $p(\boldsymbol{x})$however. This trouble can be ridden out generating a sequence of random vectors $\{\boldsymbol{x}_t\}$ via any random process such that the frequencies' distribution of sampled $\boldsymbol{x}_t$s approximates $p(\boldsymbol{x})$. Let $\{\boldsymbol{x}_0, \boldsymbol{x}_1, ...\}$ be a sequence of random vectors, and assume that at each time $t \geq 0$ $\boldsymbol{x}_{t+1}$ is sampled from a generic pdf $\pi(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t)$. Note that $\pi(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t)$ depends on the current *state* $\boldsymbol{x}_t$ but is independent of the previous states $\{\boldsymbol{x}_0, \boldsymbol{x}_1, ..., \boldsymbol{x}_{t-1}\}$ $\forall t$. Such a sequence is named *Markov chain*. Furthermore, let us assume that the Markov chain is *time homogeneus*, i.e. $\pi(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t) = \pi(\boldsymbol{x}_{t'+1}|\boldsymbol{x}_{t'})$, $\forall t \neq t'$, $t' = 1, ..., R$. Subject to regularity

conditions (see Gilks et al. 1998, Chapter 4) which we will assume to hold in the remainder of the text, as t increases the chain converges to a *stationary* distribution $\varphi(\cdot)$, which is unique and does not depend neither on t nor on the initial state $\boldsymbol{x}_0$. This implies that after an appropriate number q of iterations called *burn-in*, $\{\boldsymbol{x}_t\}$ will be approximatively sampled from $\varphi(\cdot)$, $t = q+1, ..., R$. Thus, if we set up a Markov chain sucht that its stationary distribution $\varphi(\cdot)$ is the pdf $p(\boldsymbol{x})$, then (4.9) may be approximated by the *ergodic average*

$$\widetilde{E}[\vartheta(\boldsymbol{x})] = \frac{1}{m}\sum_{t=q+1}^{R}\vartheta(\boldsymbol{x}_t), \qquad m = R - q. \tag{4.11}$$

4.2.2 MCEM algorithm

Let us assume we aim at maximizing the incomplete data log-likelihood (4.2) by the EM algorithm described in Section 4.1, but that the conditional expectation of the complete data log-likelihood function

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^h) = E_{\boldsymbol{\theta}^h}[\ell_c(\boldsymbol{\theta}|\boldsymbol{x})|\boldsymbol{y}] = \int_{\mathscr{X}(\boldsymbol{y})}\ell_c(\boldsymbol{\theta}|\boldsymbol{x})dp(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}^h)$$

expressed by (4.3) is not analitically solvable. It is straightforward to link $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^h)$ to the expected value (4.9), by setting $\vartheta(\boldsymbol{x}) = \ell_c(\boldsymbol{\theta}|\boldsymbol{x})$ and $p(\boldsymbol{x}) = p(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}^h)$. Hence, we may approximate the E-step (4.3) by

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^h) = \frac{1}{R}\sum_{t=1}^{R}\ell_c(\boldsymbol{\theta}|\boldsymbol{x}_t), \tag{4.12}$$

as in (4.10). Otherwise, if we are not able to sample m independent values of $\boldsymbol{x}$ from p$(\boldsymbol{x}|\boldsymbol{y})$ directly, we may still approximate (4.3) by

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^h) = \frac{1}{m}\sum_{t=q+1}^{R}\ell_c(\boldsymbol{\theta}|\boldsymbol{x}_t), \tag{4.13}$$

which is the ergodic average (4.11). The EM algorithm in Section 4.1 becomes an MCEM algorithm, that at iteration $h+1$ is given by the following steps:

1. Draw m samples of $\boldsymbol{x}$ from $p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta}^h)$,

2. E-step: compute $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^h)$ as in (4.12) or (4.13),

3. M-step: compute $\boldsymbol{\theta}^{h+1} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^h)$ and set $\boldsymbol{\theta}^{h+1} = \boldsymbol{\theta}^h$.

4.3 Gibbs sampler

In order to approximate (4.9) by (4.11) we must build up a Markov chain such that its stationary distribution $\varphi(\cdot)$ converges to $p(\boldsymbol{x})$. A method to achieve this objective is the Gibbs sampler (Gelfand and Smith, 1990; Geman and Geman, 1984), which is applied in the remainder of this work. In order to explain the Gibbs sampler we need first to introduce the Metropolis-Hastings algorithm (MH), a generalization presented by Hastings (1970) of the algorithm proposed by Metropolis et al. (1953). Let $g(\cdot|\boldsymbol{x}_t)$ be a *proposal distribution* from which we sample the *candidate* $\boldsymbol{x}^*$ for the next state $\boldsymbol{x}_{t+1}$. Then, given the realization $\boldsymbol{x}_t$, the $(t+1)$th iteration of MH is:

1. Sample $\boldsymbol{x}^*$ from $g(\cdot|\boldsymbol{x}_t)$,

2. Compute the *acceptance* probability

$$\alpha(\boldsymbol{x}_t, \boldsymbol{x}^*) = \min\left[1, \frac{p(\boldsymbol{x}^*)g(\boldsymbol{x}_t|\boldsymbol{x}^*)}{p(\boldsymbol{x}_t)g(\boldsymbol{x}^*|\boldsymbol{x}_t)}\right],$$

3. Sample a value u from a uniform U(0,1),

4. If $u \leq \alpha(\boldsymbol{x}_t, u)$ set $\boldsymbol{x}_{t+1} = \boldsymbol{x}^*$, $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t$ otherwise,

5. Set $t = t + 1$ and start again from step 1.

Note that after an appropriate burn-in period $\boldsymbol{x}_t$ is approximatively sampled from $p(\cdot)$, whatever the proposal distribution is (see Gilks et al. 1998, p. 7).

The MH may also be implemented by partitioning the random vector $\boldsymbol{x}_{\cdot}$ into k components, $\boldsymbol{x}_{\cdot} = [\boldsymbol{x}_{\cdot,1}, \boldsymbol{x}_{\cdot,2}, ..., \boldsymbol{x}_{\cdot,c}, ..., \boldsymbol{x}_{\cdot,k}]$, then updating such components sequentially, i.e each iteration is divided into k updating steps. The generic component $\boldsymbol{x}_{\cdot,c}$ could be either a vector or a scalar, and the components may also be of different dimensions. Let $\boldsymbol{x}_{\cdot,-c} = [\boldsymbol{x}_{\cdot,1},...\boldsymbol{x}_{\cdot,c-1}, ..., \boldsymbol{x}_{\cdot,c+1}, ..., \boldsymbol{x}_{\cdot,k}]$ be the vector without the component $\boldsymbol{x}_{\cdot,c}$. Then, at the iteration $h + 1$, the component $\boldsymbol{x}_{t,c}$ is updated by generating a candidate $\boldsymbol{x}^*_{\cdot,c}$ which is sampled from the proposal pdf $g_c(\cdot|\boldsymbol{x}_{t,c}, \boldsymbol{x}_{t,-c})$, where $\boldsymbol{x}_{t,-c} =[\boldsymbol{x}_{t+1,1},...,\boldsymbol{x}_{t+1,c-1},...,\boldsymbol{x}_{t,c+1},...,\boldsymbol{x}_{t,k}]$, the $c-1$ components having been updated already. The acceptance probability of $\boldsymbol{x}^*_{\cdot,c}$ is given by

$$\alpha(\boldsymbol{x}_{t,-c}, \boldsymbol{x}_{t,c}, \boldsymbol{x}^*_{\cdot,c}) = \min\left[1, \frac{p(\boldsymbol{x}^*_{\cdot,c}|\boldsymbol{x}_{t,-c})g_c(\boldsymbol{x}_{t,c}|\boldsymbol{x}^*_{\cdot,c}, \boldsymbol{x}_{t,-c})}{p(\boldsymbol{x}_{t,c}|\boldsymbol{x}_{t,-c})g_c(\boldsymbol{x}^*_{\cdot,c}|\boldsymbol{x}_{t,c}, \boldsymbol{x}_{t,-c})}\right], \tag{4.14}$$

where $p(\boldsymbol{x}_{\cdot,c}|\boldsymbol{x}_{\cdot,-c})$ is a *full conditional* distribution, that is the pdf of the cth component of $\boldsymbol{X}$ given the realizations of all the other components.

The Gibbs sampler is a particular case of the partitioned MH previously exposed, in which we set

$$g_c(\cdot|\boldsymbol{x}_{t,c}, \boldsymbol{x}_{t,-c}) = p(\cdot|\boldsymbol{x}_{\cdot,-c}) \tag{4.15}$$

that is the proposal distribution of the candidate $\boldsymbol{x}^*_{\cdot,c}$ is given by its full conditional distribution. Note that the acceptance probability of Gibbs sampler is always equal to 1 and the candidates are therefore always accepted because

from (4.14) it is

$$\alpha(\boldsymbol{x}_{t,-c}, \boldsymbol{x}_{t,c}, \boldsymbol{x}^*_{\cdot,c}) = \min\left[1, \frac{p(\boldsymbol{x}^*_{\cdot,c}|\boldsymbol{x}_{t,-c})p(\boldsymbol{x}_{t,c}|\boldsymbol{x}_{t,-c})}{p(\boldsymbol{x}_{t,c}|\boldsymbol{x}_{t,-c})p(\boldsymbol{x}^*_{\cdot,c}|\boldsymbol{x}_{t,-c})}\right] = 1,$$

setting $g_c(\boldsymbol{x}^*_{\cdot,c}|\boldsymbol{x}_{t,c}, \boldsymbol{x}_{t,-c}) = p(\boldsymbol{x}^*_{\cdot,c}|\boldsymbol{x}_{t,-c})$ and $g_c(\boldsymbol{x}_{t,c}|\boldsymbol{x}^*_{\cdot,c}, \boldsymbol{x}_{t,-c}) = p(\boldsymbol{x}_{t,c}|\boldsymbol{x}_{t,-c})$ respectively because of (4.15). The Gibbs sampler consists of sampling from full conditional distributions, that is at the $(t+1)$th iteration it is given by

1. Sample $\boldsymbol{x}_{t+1,1}$ from $p(\boldsymbol{x}^*_{\cdot 1}|\boldsymbol{x}_{t,2}, ..., \boldsymbol{x}_{t,c}, ..., \boldsymbol{x}_{t,\cdot k})$

2. Sample $\boldsymbol{x}_{t+1,2}$ from $p(\boldsymbol{x}^*_{\cdot 2}|\boldsymbol{x}_{t+1,1}, \boldsymbol{x}_{t,3}, ..., \boldsymbol{x}_{t,c}, ..., \boldsymbol{x}_{t,\cdot k})$

$\vdots$

k. Sample $\boldsymbol{x}_{t+1,k}$ from $p(\boldsymbol{x}^*_{\cdot k}|\boldsymbol{x}_{t+1,1}, ..., \boldsymbol{x}_{t+1,c}, ..., \boldsymbol{x}_{t,k-1})$

k+1. Set $t = t + 1$ then start again from step 1.

4.3.1 Adaptive Rejection Sampling

Without loss of generality, let us assume that the generic vector $\boldsymbol{x}_\cdot$ can be parted into k *scalar* components, i.e. $\boldsymbol{x}_\cdot = [x_{\cdot,1}, x_{\cdot,2}, ..., x_{\cdot,c}, ..., x_{\cdot,k}]'$. We may write the cth full conditional distribution as $p(x_{\cdot,c}|\boldsymbol{x}_{\cdot,-c})\,\forall c$ and the Gibbs sampler described in the previous section would be given by

1. Sample $x_{t+1,1}$ from $p(x^*_{\cdot 1}|x_{t,2}, ..., x_{t,c}, ..., x_{t,\cdot k})$

2. Sample $x_{t+1,2}$ from $p(x^*_{\cdot 2}|x_{t+1,1}, x_{t,3}, ..., x_{t,c}, ..., x_{t,\cdot k})$

$\vdots$

k. Sample $x_{t+1,k}$ from $p(x^*_{\cdot k}|x_{t+1,1}, ..., x_{t+1,c}, ..., x_{t,k-1})$

k+1. Set $t = t + 1$ then start again from step 1.

If we assume that $p(x_{.,c}|\boldsymbol{x}_{.,-c})$ is a density function $\forall c$ and that *each* univariate full conditional density is *log-concave*, an efficient way to sample from them is the Adaptive Rejection Sampling (ARS) proposed by Gilks and Wild (1992). Such sampling procedure is a variant of the standard rejection sampling which we briefly describe.

Our purpose is to sample a value $x_{t+1,c}\,\forall c$ from $p(x_{.,c}|\boldsymbol{x}_{t,-c})$ which we define as our *target* distribution. If $p(x_{.,c}|\boldsymbol{x}_{t,-c})$ is not a standard function the computation of a normalizing constant is needed. The rejection sampling allows us to skip such computation. This is a great advantage because it may happen that the normalizing constant for full conditional distributions has not a closed form. Let $g(\cdot)$ be a generic density function such that $g(x_{.,c}) \propto p(x_{.,c}|\boldsymbol{x}_{t,-c})$, $\forall x, c$. Let also $G(\cdot)$ be an arbitrary *envelope* function of $g(\cdot)$ such that it is $G(x_{.,c}) > g(x_{.,c})\,\forall x, c$.

Let $x^*_{.c}$ be the *candidate* point for $x_{t+1,c}$. For each cth component, the rejection sampling (RS) consists in the following algorithm:

1. Sample a point $x^*_{.,c}$ from $G(x_{.,c})$;

2. Sample a point u from a Uniform$(0, 1)$;

3. If $u \leq g(x^*_{.,c})/G(x^*_{.,c})$ accept $x^*_{.,c}$ and set $x^*_{.,c} = x_{t+1,c}$, if else reject it;

4. Iterate until a candidate $x^*_{.,c}$ is accepted.

Note that we do not need to evaluate the normalizing constant for $p(x_{.,c}|\boldsymbol{x}_{t,-c})$. This is an important advantage of the rejection sampling as generally the normalizing constant for full conditional distributions has not a closed form. In

order to minimize the number of possible rejections the envelope function $G(\cdot)$ must be as closer as possible to $g(x^*_{.,c})$, but this can be difficult because we need to find the maximum of $g(\cdot)$ in its dominion D. This may require several computations of $g(\cdot)$. If we assume that $p(x_{.,c}|\boldsymbol{x}_{t,-c})$ belongs to the class of log-concave densities $\forall c$, then $\dot{g()}$ is log-concave as well. The Adaptive Rejection Sampling consists then in building up an envelope function $\log G(\cdot)$ which is a *piecewise linear* upper hull formed by the tangents to $\log g(\cdot)$ in few set of points $S \subset D$. The adaptivity of the algorithm is given by the fact that if a candidate $x^*_{.,c}$ is rejected it is included in the set S tightening the envelope and reducing the probability of rejection of the next candidate. The Adaptive Rejection Sampling is then given by

1. Sample a point $x^*_{.,c}$ from $G(x_{.,c})$;

2. Sample a point u from a Uniform$(0, 1)$;

3. If $u \leq g(x^*_{.,c})/G(x^*_{.,c})$ accept $x^*_{.,c}$ and set $x^*_{.,c} = x_{t+1,c}$, if else reject it;

4. Include $x^*_{.,c}$ in S

5. Iterate until a candidate $x^*_{.,c}$ is accepted.

The Adaptive Rejection Sampling reduces the number of evaluations of $\log g(\cdot)$ because the log-concavity assumption let us skip the need to ocate the maximum of $g(\cdot)$. Furthermore, the increasing tightening of the envelope $\log G(\cdot)$ reduces the probability of rejection of the next candidate, i.e. it reduces the probability of further computations of $g(\cdot)$.

Chapter 5

Missing values

It is frequent in survival studies that some subject can have missing values of one or more covariates. There are three popular approaches to deal with this problem:

- The observational units presenting missing values are simply dropped (complete case analysis).

- Missing values can be replaced by imputed ones according to some procedure based on complete cases.

- The data set can be left as it is and handled by the EM algorithm.

Whatever choice is made attention is needed, because an incautious handling of missing values could distort the analysis results. Three things must be considered in the choice of the method to apply: the missing data pattern, the underlying missing values generating process, the purpose of the analysis. According to Rubin (1976), missing values must be regarded as a probabilistic phenomenon. Three possible kinds of missing values generating processes exist. Let us say we are studying life times of n subjects and that for each

ith subject there are three random variables (RV):

- X_i : age
- F_i : number of cigarettes smoked per day
- R_i : a missingness RV which is 1 if X_i or F_i is missing and 0 if it is observed.

Let us also suppose that age is completely observed, while the number of smoked cigarettes has some unobserved values. We could face one of the following situations:

1. The probability of R=1 is independent of both age and cigarettes number. In this case missing values are *Missing Completely At Random(MCAR),* $P(R_i = 1|x_i, f_i) = P(R = 1)$.

2. *The probability of R=1 depends on the value of age but not on the number of cigarettes. In this case missing data are Missing At Random (MAR),* $P(R = 1|x_i, f_i) = P(R = 1|x_i)$.

3. *The probability of R=1 depends both on cigarettes number and age. The data are Missing Not At Random (MNAR).*

The complete case analysis provides biased estimates of the parameters of interest unless the data are MCAR or the sample size is sufficiently large. If the missing data are MAR both the imputation method and the EM algorithm are valid. However, by the imputation method a "likely" value is

imputed replacing each missing one, then the usual estimation procedures are applied. In this way the analysis is performed treating the imputed values as if they had been observed. In the EM approach, for each missing value a set of all (or sampled) possible values is considered, so that the distinction between what has been observed and what has been not still holds. A more detailed review about the missingness generating processes and the possible consequences of ignoring them can be found in Little and Rubin (1987) and Schafer and Graham (2002).

5.1 PH model with missing values

Without loss of generality, let us assume we want to study the failure times $\{t_i\}$ of n observed individuals belonging to an *homogeneus* population, where T is a positive random variable (RV) from which failure times t_i are drawn, $i = 1, ...n$. Let also define the positive RVs C and W, from which respectively censoring times c_i and entering times w_i are independently drawn. Because of possible left-truncation and right-censoring we observe the event time y which is the realization of the RV $Y = \min(T, C)$. Let also δ_i be a failure indicator which equals 1 if a failure occurs, 0 otherwise, i.e. if there is an event or a censoring at time y_i. Let $\boldsymbol{x}_i = [x_{i1}, ..., x_{ik}]'$ be a k-vector of covariates, assuming that $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are the realizations of n independent RVs $\boldsymbol{X}_i \sim p_{\boldsymbol{x}}(\boldsymbol{x}_i|\boldsymbol{\alpha})$, $\boldsymbol{\alpha}$ a vector of parameters. Each ith subject is therefore described by $(y_i, \delta_i, \boldsymbol{x}_i, w_i)$ (see Chapter 2). Along with (Herring and Ibrahim, 2001) and recalling (2.6), if $\boldsymbol{x}_i$ is *fully observed* we may write the ith subject's likelihood contribution as

$$
\begin{aligned}
&L_i^c(\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\alpha} | y_i, \boldsymbol{x}_i, \delta_i, w_i) = \\
&= \left[\frac{p_y(y_i|\boldsymbol{x}_i, \boldsymbol{\theta}) S_c(y_i|\boldsymbol{x}_i, \boldsymbol{\psi})}{S_t(w_i|\boldsymbol{x}_i, \boldsymbol{\theta})}\right]^{\delta_i} \left[\frac{S_t(y_i|\boldsymbol{x}_i, \boldsymbol{\theta}) p_c(y_i|\boldsymbol{x}_i, \boldsymbol{\psi})}{S_t(w_i|\boldsymbol{x}_i, \boldsymbol{\theta})}\right]^{1-\delta_i} \times p_{\boldsymbol{x}}(\boldsymbol{x}_i|\boldsymbol{\alpha}) \\
&= [p_y(y_i|\boldsymbol{x}_i, \boldsymbol{\theta})]^{\delta} [S_t(y_i|\boldsymbol{x}_i, \boldsymbol{\theta})]^{1-\delta} \times [p_c(y_i|\boldsymbol{x}_i, \boldsymbol{\psi})]^{1-\delta_i} [S_c(y_i|\boldsymbol{x}_i, \boldsymbol{\psi})]^{\delta_i} \\
&\times \quad S_t(w_i|\boldsymbol{x}_i, \boldsymbol{\theta})^{-1} \times p_{\boldsymbol{x}}(\boldsymbol{x}_i|\boldsymbol{\alpha}) \\
&= [h_t(y_i|\boldsymbol{x}_i, \boldsymbol{\theta})]^{\delta_i} S_t(y_i|\boldsymbol{x}_i, \boldsymbol{\theta}) [S_t(w_i|\boldsymbol{x}_i, \boldsymbol{\theta})]^{-1} \\
&\times \quad [h_c(y_i|\boldsymbol{x}_i, \boldsymbol{\psi})]^{1-\delta_i} S_c(y_i|\boldsymbol{x}_i, \boldsymbol{\psi}) \times p(\boldsymbol{x}_i|\boldsymbol{\alpha}), \qquad \boldsymbol{\theta} \neq \boldsymbol{\alpha}, \qquad (5.1)
\end{aligned}
$$

where $p_y(\cdot|\boldsymbol{x}_i, \boldsymbol{\theta})$ and $S_t(\cdot|\boldsymbol{x}_i, \boldsymbol{\theta})$ are the pdf and the survival function of exit time of subject i. Analougously, $p_c(\cdot|\boldsymbol{x}_i, \boldsymbol{\psi})$ and $S_c(\cdot|\boldsymbol{x}_i, \boldsymbol{\psi})$ are the subject-specific conditional pdf and survival function of censoring time, while $p(\cdot|\boldsymbol{\alpha})$ is the pdf of $\boldsymbol{x}_i$.

Now let us assume that some of the n subjects miss the values for one or more of their covariates. Both the covariates with missing and their number-can be subject-specific. Without loss in generality we may therefore denote the ith covariates' profile as $\boldsymbol{x}_i = [\boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i}]'$, where $\boldsymbol{x}_{obs,i}$ and $\boldsymbol{x}_{mis,i}$ denote the observed and missing components of $\boldsymbol{x}_i$ respectively. Let us also define a vector $\mathbf{r}_i$ whose component r_{ij} is 1 if x_{ij} is observed, 0 otherwise, $j = 0, 1, ..., k$. The vector $\boldsymbol{r}_i$ denotes the individual missingness profile of the ith subject with pdf $p_r(\boldsymbol{r}_i|y_i, \boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i}, \delta_i, w_i, \boldsymbol{\phi})$, $\boldsymbol{\phi}$ a vector of parameters. For example, if it is $k = 3$ and $\boldsymbol{x}_i = [x_{obs,i,1}, x_{mis,i,2}, x_{obs,i,3}]'$, it is $\boldsymbol{r}_i = [1, 0, 1]'$. Hence, when missing values are present the *incomplete data* likelihood contribution of the ith subject is given by

$$L_i(\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\phi}, \boldsymbol{\alpha} | y_i, \boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i}, \delta_i, w_i, \boldsymbol{r}_i) =$$

$$= \int_{\mathscr{X}_i} \Big\{ p_r(\boldsymbol{r}_i | y_i, \boldsymbol{x}_i, \delta_i, w_i, \boldsymbol{\phi}) \times [h_t(y_i|\boldsymbol{x}_i, \boldsymbol{\theta})]^{\delta_i} S_t(y_i|\boldsymbol{x}_i, \boldsymbol{\theta}) [S_t(w_i|\boldsymbol{x}_i, \boldsymbol{\theta})]^{-1}$$
$$\times [h_c(y_i|\boldsymbol{x}_i, \boldsymbol{\psi})]^{1-\delta_i} S_c(y_i|\boldsymbol{x}_i, \boldsymbol{\psi}) \Big\} dp(\boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i} | \boldsymbol{\alpha}), \tag{5.2}$$

where $\boldsymbol{x}_i = [\boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i}]'$ and $\mathscr{X}_i$ is the set of all possible *complete data patterns* $\boldsymbol{x}_i$ associated to $\boldsymbol{x}_{obs,i}$. In order to clarify this concept let us assume again $\boldsymbol{x}_i = [x_{i,1}, x_{i,2}, x_{i,3}]', \forall i$, where x_2 and $x_{3,}$ are binary. Let also $x_{2,i}, x_{3,i}$ be missing for the ith subject only, such that his covariate profile is given by $\boldsymbol{x}_i = [\boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i}]' = [x_{obs,i,1}, x_{mis,i,2}, x_{mis,i,3}]'$. The set of all possible complete data patterns $\mathscr{X}_i$ would then be

$$\mathscr{X}_i = \{[x_{obs,i,1}, 0, 0]', [x_{obs,i,1}, 0, 1]', [x_{obs,i,1}, 1, 0]', [x_{obs,i,1}, 1, 1]'\}\,.$$

Hence, in the current notation $p(\boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i}|\boldsymbol{\alpha})$ denotes the prior distribution of *each* possible complete data pattern $\boldsymbol{x}_i = [\boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i}]' \in \mathscr{X}_i$, in our example

$$\begin{aligned} p(\boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i}|\boldsymbol{\alpha}) \;\; = \;\; & \{p(x_{obs,i,1}, 0, 0|\boldsymbol{\alpha}), p(x_{obs,i,1}, 0, 1|\boldsymbol{\alpha}), \\ & p(x_{obs,i,1}, 1, 0|\boldsymbol{\alpha}), p(x_{obs,i,1}, 1, 1|\boldsymbol{\alpha})\}. \end{aligned}$$

Note that in this example it is $\boldsymbol{r}_i = [1, 0, 0]'$ and $p_r(\boldsymbol{r}_i|y_i, \boldsymbol{x}_i, \delta_i, w_i, \boldsymbol{\phi})$ is the joint pdf of the missing data pattern $\boldsymbol{x}_i = [x_{obs,i,1}, x_{mis,i,2}, x_{mis,i,3}]'$.
Let us assume that:

1. Missing values are MAR, that is

$$p_r(\boldsymbol{r}_i|y_i, \boldsymbol{x}_i, \delta_i, w_i, \boldsymbol{\phi}) = p_r(\boldsymbol{r}_i|y_i, \boldsymbol{x}_{obs,i}, \delta_i, w_i, \boldsymbol{\phi}).$$

2. The censoring distribution is independent of missing values, which implies

$$h_c(y_i|\boldsymbol{x}_i, \boldsymbol{\psi})^{1-\delta_i} S_c(y_i|\boldsymbol{x}_i, \boldsymbol{\psi}) = h_c(y_i|\boldsymbol{x}_{obs,i}, \boldsymbol{\psi})^{1-\delta_i} S_c(y_i|\boldsymbol{x}_{obs,i}, \boldsymbol{\psi}).$$

3. $\boldsymbol{\theta}, \boldsymbol{\alpha} \neq \boldsymbol{\phi}, \boldsymbol{\psi}$.

Because of assumptions 1. e 2. we may write (5.2) as

$$L_i(\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\phi}, \boldsymbol{\alpha} | y_i, \boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i}, \delta_i, w_i, \boldsymbol{r}_i) =$$

$$= \; p_r(\boldsymbol{r}_i | y_i, \boldsymbol{x}_{obs,i}, \delta_i, w_i, \boldsymbol{\phi}) \times h_c(y_i | \boldsymbol{x}_{obs,i}, \boldsymbol{\psi})^{1-\delta_i} S_c(y_i | \boldsymbol{x}_{obs,i}, \boldsymbol{\psi})$$

$$\times \; \int_{\mathscr{X}_i} \Big\{ [h_t(y_i | \boldsymbol{x}_i, \boldsymbol{\theta})]^{\delta_i} S_t(y_i | \boldsymbol{x}_i, \boldsymbol{\theta}) \times [S_t(w_i | \boldsymbol{x}_i, \boldsymbol{\theta})]^{-1} \Big\} dp(\boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i} | \boldsymbol{\alpha}).$$

Assumption 3. allows us to write the ith incomplete data likelihood distribution (5.2) as

$$L_i(\boldsymbol{\theta}, \boldsymbol{\alpha} | y_i, \boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i}, \delta_i, w_i) =$$

$$\propto \int_{\mathscr{X}_i} \Big\{ [h_t(y_i | \boldsymbol{x}_i, \boldsymbol{\theta})]^{\delta_i} S_t(y_i | \boldsymbol{x}_i, \boldsymbol{\theta}) \times [S_t(w_i | \boldsymbol{x}_i, \boldsymbol{\theta})]^{-1} \Big\} dp(\boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i} | \boldsymbol{\alpha}). \tag{5.3}$$

5.2 Categorical missing covariates

Starting from (5.3) and assuming that all of the components of $\boldsymbol{x}_{mis,i}$ are discrete $\forall i$, the ith incomplete data log-likelihood contribution may be written as

$$\ell_i(\boldsymbol{\xi} | y_i, \boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i}, \delta_i, w_i) =$$

$$= \log \sum_{\mathscr{X}_i} \Big\{ [h_t(y_i | \boldsymbol{x}_i, \boldsymbol{\theta})]^{\delta_i} S_t(y_i | \boldsymbol{x}_i, \boldsymbol{\theta}) \times [S_t(w_i | \boldsymbol{x}_i, \boldsymbol{\theta})]^{-1} \Big\} \times p(\boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i} | \boldsymbol{\alpha}). \tag{5.4}$$

where $\boldsymbol{\xi} = [\boldsymbol{\theta}, \boldsymbol{\alpha}]'$. In order to estimate $\boldsymbol{\xi}$ we may maximize the incomplete data log-likelihood $\ell(\boldsymbol{\xi} | \boldsymbol{y}, \boldsymbol{X}_{mis}, \boldsymbol{X}_{obs}, \boldsymbol{\delta}, \boldsymbol{w}) = \sum_i \ell_i(\boldsymbol{\xi} | y_i, \boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i}, \delta_i, w_i)$, where $\boldsymbol{X}_{mis} = [\boldsymbol{x}_{mis,1}, ..., \boldsymbol{x}_{mis,n}]'$ and $\boldsymbol{X}_{obs} = [\boldsymbol{x}_{obs,1}, ..., \boldsymbol{x}_{obs,n}]'$.

If such maximization is not possible, along with Lipsitz and Ibrahim (1996b) we may exploit the EM algorithm described in Chapter 4 to find MLEs $\hat{\boldsymbol{\xi}}$, i.e. by iteratively maximizing the *expected complete data* log-likelihood given the observed data defined by (4.3). In the current setting the ith contribution to it is given by

$$E[\ell_i^c(\boldsymbol{\xi}|y_i, \boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i}, \delta_i, w_i)|y_i, \boldsymbol{x}_{obs,i}, \delta_i, w_i] =$$

$$= \sum_{\mathscr{X}_i} p(\boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i}|y_i, \boldsymbol{x}_{obs,i}, \delta_i, w_i, \boldsymbol{\xi}) \times \ell_i^c(\boldsymbol{\xi}|y_i, \boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i}, \delta_i, w_i), \quad (5.5)$$

where

$$\ell_i^c(\boldsymbol{\xi}|y_i, \boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i}, \delta_i, w_i) =$$

$$= \log\left\{[h_t(y_i|\boldsymbol{x}_i, \boldsymbol{\theta})]^{\delta_i} S_t(y_i|\boldsymbol{x}_i, \boldsymbol{\theta})[S_t(w_i|\boldsymbol{x}_i, \boldsymbol{\theta})]^{-1} \times p(\boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i}|\boldsymbol{\alpha})\right\} \quad (5.6)$$

is the *complete data* log-likelihood contribution associated to each possible complete data pattern $(\boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i})$ weighted by its *posterior* probability

$$p(\boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i}|y_i, \boldsymbol{x}_{obs,i}, \delta_i, w_i, \boldsymbol{\xi}).$$

Note that usually the interest lies in estimating $\boldsymbol{\theta}$, that is $\boldsymbol{\alpha}$ is considered a vector of nuisance parameters. We should therefore aim at minimizing the number of such nuisance parameters in order to reduce the computational burden required for the estimation procedure. Along with Lipsitz and Ibrahim (1996a), the pdf of $\boldsymbol{x}_i$ may be modeled as a product of one-dimensional full conditional distributions,

$$\begin{aligned} p(\boldsymbol{x}_i|\boldsymbol{\alpha}) &= p(x_{i,k}|x_{i,1}, x_{i,2}, ..., x_{i,k-1}|\boldsymbol{\alpha}_k) \\ &\times p(x_{i,k-1}|x_{i,1}, x_{i,2}, ..., x_{i,k-2}|\boldsymbol{\alpha}_{k-1}) \cdots \times p(x_{i,1}|\boldsymbol{\alpha}_1) \end{aligned} \quad (5.7)$$

where $\boldsymbol{\alpha}_j$ is a vector of parameters for the j^{th} conditional distribution, $j = 1, ..., k$, $\boldsymbol{\alpha}_j \neq \boldsymbol{\alpha}_{j'}$, $\forall j \neq j', j' = 1, ..., k$, and $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, ...\boldsymbol{\alpha}_k)$. Without loss of generality, if there are missing values let us assume that the components with missing values are the last $k - r$ of $\boldsymbol{x}_i$, $0 < r < k$.

Hence, we may write the pdf of the generic complete data pattern as

$$\begin{aligned} p(\boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i}|\boldsymbol{\alpha}) &= p(x_{mis,i,k}|x_{mis,i,k-1}, ..., x_{mis,i,r+1}, \boldsymbol{x}_{obs,i}, \boldsymbol{\alpha}_k) \\ &\times p(x_{mis,i,k-1}|x_{mis,i,k-2}, ..., x_{mis,i,r+1}, \boldsymbol{x}_{obs,i}, \boldsymbol{\alpha}_{k-1}) \\ &\vdots \\ &\times p(x_{mis,i,r+1}|\boldsymbol{x}_{obs,i}, \boldsymbol{\alpha}_{r+1}) \\ &\times p(x_{obs,i,r}|x_{obs,i,r-1}, ..., x_{obs,i,1}, \boldsymbol{\alpha}_r) \\ &\vdots \\ &\times p(x_{obs,i,1}|\boldsymbol{\alpha}_1). \end{aligned}$$

Note that we need to specify a distribution for $\boldsymbol{x}_{mis,i}$ only. Thus, we may replace $p(\boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i}|\boldsymbol{\alpha})$ in (5.6) with the conditional probability of the *missing data pattern* $\boldsymbol{x}_{mis,i}$,

$$\begin{aligned} p(\boldsymbol{x}_{mis,i}|\boldsymbol{x}_{obs,i}, \boldsymbol{\alpha}) &= p(x_{mis,i,k}|x_{mis,i,k-1}, ..., x_{mis,i,r+1}, \boldsymbol{x}_{obs,i}, \boldsymbol{\alpha}_k) \\ &\times p(x_{mis,i,k-1}|x_{mis,i,k-2}, ..., x_{mis,i,r+1}, \boldsymbol{x}_{obs,i}, \boldsymbol{\alpha}_{k-1}) \\ &\vdots \\ &\times p(x_{mis,i,r+1}|\boldsymbol{x}_{obs,i}, \boldsymbol{\alpha}_{r+1}) \\ &\propto p(\boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i}). \end{aligned} \tag{5.8}$$

Similarly, we may define the posterior probability of $\boldsymbol{x}_{mis,i}$ as a full conditional

as well as (5.8),

$$\begin{aligned} p(\boldsymbol{x}_{mis,i}|y_i, \boldsymbol{x}_{obs,i}, \delta_i, w_i, \boldsymbol{\xi}) &= p(x_{mis,i,k}|y_i, x_{mis,i,k-1}, .., x_{mis,i,r}, \boldsymbol{x}_{obs,i}, \delta_i, w_i, \boldsymbol{\xi}^h) \\ &\times p(x_{mis,i,k-1}|y_i, x_{mis,i,k-2}, .., x_{mis,i,r}, \boldsymbol{x}_{obs,i}, \delta_i, w_i, \boldsymbol{\xi}^h) \\ &\vdots \\ &\times p(x_{mis,i,r}|y_i, \boldsymbol{x}_{obs,i}, \delta_i, w_i, \boldsymbol{\xi}^h) \qquad (5.9) \\ &\propto p(\boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i}|y_i, \boldsymbol{x}_{obs,i}, \delta_i, w_i, \boldsymbol{\xi}), \end{aligned}$$

then substitute it in (5.5). Hence, the ith contribution to the expected conditional complete data log-likelihood may be written as

$$E[\ell_i^c(\boldsymbol{\xi}|y_i, \boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i}, \delta_i, w_i)|y_i, \boldsymbol{x}_{obs,i}, \delta_i, w_i] =$$

$$= \sum_{\boldsymbol{x}_{mis,i}} p(\boldsymbol{x}_{mis,i}|y_i, \boldsymbol{x}_{obs,i}, \delta_i, w_i, \boldsymbol{\xi}) \times \ell_i^c(\boldsymbol{\xi}|y_i, \boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i}, \delta_i, w_i), \qquad (5.10)$$

where $\ell_i^c(\boldsymbol{\xi}|y_i, \boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i}, \delta_i, w_i) \propto$

$$\log\left\{[h_t(y_i|\boldsymbol{x}_i, \boldsymbol{\theta})]^{\delta_i} S_t(y_i|\boldsymbol{x}_i, \boldsymbol{\theta})[S_t(w_i|\boldsymbol{x}_i, \boldsymbol{\theta})]^{-1} \times p(\boldsymbol{x}_{mis,i}|\boldsymbol{x}_{obs,i}, \boldsymbol{\alpha})\right\}. \qquad (5.11)$$

The sum is over the possible missing data patterns of subject i, each weighted by $p(\boldsymbol{x}_{mis,i}|y_i, \boldsymbol{x}_{obs,i}, \delta_i, w_i, \boldsymbol{\xi})$ which may be written as a *posterior* probability:

$$p(\boldsymbol{x}_{mis,i}|y_i, \boldsymbol{x}_{obs,i}, \delta_i, w_i, \boldsymbol{\xi}) =$$

$$= \frac{[h_t(y_i|\boldsymbol{x}_i, \boldsymbol{\theta})]^{\delta_i} S_t(y_i|\boldsymbol{x}_i, \boldsymbol{\theta})[S_t(w_i|\boldsymbol{x}_i, \boldsymbol{\theta})]^{-1} \times p(\boldsymbol{x}_{mis,i}|\boldsymbol{x}_{obs,i}, \boldsymbol{\alpha})}{\sum_{\boldsymbol{x}_{mis,i}} [h_t(y_i|\boldsymbol{x}_i, \boldsymbol{\theta})]^{\delta_i} S_t(y_i|\boldsymbol{x}_i, \boldsymbol{\theta})[S_t(w_i|\boldsymbol{x}_i, \boldsymbol{\theta})]^{-1} \times p(\boldsymbol{x}_{mis,i}|\boldsymbol{x}_{obs,i}, \boldsymbol{\alpha})}. \qquad (5.12)$$

The weighted sum (5.10) provides the contribution of a subject with som missing covariates. If we consider a subject $i' \neq i$ with fully observed covariates suvh that $\boldsymbol{x}_{i'} = \boldsymbol{x}_{obs,i'}$, the contribution (5.10) becomes $\ell_i(\boldsymbol{\xi}|y_i, \boldsymbol{x}_{obs,i},$

$\boldsymbol{x}_{mis,i}, \delta_i, w_i)$. For example let us assume there are $n = 2$ observed subjects, each described by a vector of binary covariates $\boldsymbol{x}_i = [x_{1,i}, x_{2,i}]'$, and also that $\boldsymbol{x}_1$ is fully observed while the first component of $\boldsymbol{x}_2$ is missing, $\boldsymbol{x}_2 = [x_{mis,1,2}, x_{obs,2,2}]'$. The individual contributions to the expected complete data log-likelihood would therefore be given by

Subject 1: $\ell_i^c(\boldsymbol{\xi}|y_1, \boldsymbol{x}_1, \delta_1, w_1)$

$$\text{Subject 2}: \begin{cases} p(x_{mis,2,1} = 0|y_2, x_{obs,2,2}, \delta_2, w_2, \boldsymbol{\xi}) \times \ell_2^c(\boldsymbol{\xi}|y_2, 0, x_{obs,2,2}, \delta_2, w_2) \\ \\ p(x_{mis,2,1} = 1|y_2, x_{obs,2,2}, \delta_2, w_2, \boldsymbol{\xi}) \times \ell_2^c(\boldsymbol{\xi}|y_2, 1, x_{obs,2,2}, \delta_2, w_2) \end{cases}$$

and the corresponding whole log-likelihood would be

$$\begin{aligned} & E[\ell(\boldsymbol{\xi}|\boldsymbol{y}, \boldsymbol{X}_{mis}, \boldsymbol{X}_{obs}, \boldsymbol{\delta}, \boldsymbol{w})|\boldsymbol{y}, \boldsymbol{X}_{obs,i}, \boldsymbol{\delta}, \boldsymbol{w}] = \\ & \quad = \ \ell_1^c(\boldsymbol{\xi}|y_1, \boldsymbol{x}_1, \delta_1, w_1) \\ & \quad + \ p(x_{mis,2,1} = 0|y_2, x_{obs,2,2}, \delta_2, w_2, \boldsymbol{\xi}) \times \ell_2^c(\boldsymbol{\xi}|y_2, 0, x_{obs,2,2}, \delta_2, w_2) \\ & \quad + \ p(x_{mis,2,1} = 1|y_2, x_{obs,2,2}, \delta_2, w_2, \boldsymbol{\xi}) \times \ell_2^c(\boldsymbol{\xi}|y_2, 1, x_{obs,2,2}, \delta_2, w_2). \end{aligned}$$

Given a set of starting values $\boldsymbol{\xi}_0 = [\boldsymbol{\theta}_0, \boldsymbol{\alpha}_0]$, at the $(h+1)$th iteration the EM algorithm is given by the following steps:

1. Compute the expected complete data log-likelihood

$$\begin{aligned} Q(\boldsymbol{\xi}|\boldsymbol{\xi}^h) &= \sum_{i=1}^{n} E[\ell_i^c(\boldsymbol{\xi}|y_i, \boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i}, \delta_i, w_i)|y_i, \boldsymbol{x}_{obs,i}, \delta_i, w_i] \\ &= \sum_{i=1}^{n} \sum_{\boldsymbol{x}_{mis,i}} \Big\{ p(\boldsymbol{x}_{mis,i}|y_i, \boldsymbol{x}_{obs,i}, \delta_i, w_i, \boldsymbol{\xi}^h) \\ &\qquad \times \ell_i^c(\boldsymbol{\xi}|y_i, \boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i}, \delta_i, w_i, \boldsymbol{\xi}^h) \Big\}, \end{aligned}$$

2. M-step: find $\boldsymbol{\theta}^{h+1} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^h)$ and set $\boldsymbol{\theta}^{h+1} = \boldsymbol{\theta}^h$,

where $p(\boldsymbol{x}_{mis,i}|y_i, \boldsymbol{x}_{obs,i}, \delta_i, w_i, \boldsymbol{\xi}^h)$ and $\ell_i^c(\boldsymbol{\xi}|y_i, \boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i}, \delta_i, w_i, \boldsymbol{\xi}^h)$ are given by (5.12) and (5.11) respectively, given the current estimates $\boldsymbol{\xi}^h = [\boldsymbol{\theta}^h, \boldsymbol{\alpha}^h]$.

5.3 Continuous missing covariates

Now let us assume that the missing covariates $\boldsymbol{x}_i$ are continuous $\forall i$. This implies that both $p(\boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i}|\boldsymbol{\alpha})$ in (5.5) and $p(\boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i}|y_i, \boldsymbol{x}_{obs,i}, \delta_i, w_i, \boldsymbol{\xi})$ in (5.6) are now density functions. Let us also assume that such densities can be expressed as products of univariate full conditional distributions, as in (5.8) and (5.9) respectively. Thus, similarly to (5.10), the ith subject's contribution may be written as

$$E[\ell_i^c(\boldsymbol{\xi}|y_i, \boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i}, \delta_i, w_i)|y_i, \boldsymbol{x}_{obs,i}, \delta_i, w_i] =$$

$$= \int_{\boldsymbol{x}_{mis,i}} p(\boldsymbol{x}_{mis,i}|y_i, \boldsymbol{x}_{obs,i}, \delta_i, w_i, \boldsymbol{\xi}) \times \ell_i^c(\boldsymbol{\xi}|y_i, \boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i}, \delta_i, w_i) d\boldsymbol{x}_{mis,i},$$

where

$$\ell_i^c(\boldsymbol{\xi}|y_i, \boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i}, \delta_i, w_i)$$

$$= [h_t(y_i|\boldsymbol{x}_i, \boldsymbol{\theta})]^{\delta_i} S_t(y_i|\boldsymbol{x}_i, \boldsymbol{\theta})][S_t(w_i|\boldsymbol{x}_i, \boldsymbol{\theta})]^{-1} \times p(\boldsymbol{x}_{mis,i}|\boldsymbol{x}_{obs,i}, \boldsymbol{\alpha}) \tag{5.13}$$

and

$$p(\boldsymbol{x}_{mis,i}|y_i, \boldsymbol{x}_{obs,i}, \delta_i, w_i, \boldsymbol{\xi}) =$$

$$= \frac{[h_t(y_i|\boldsymbol{x}_i, \boldsymbol{\theta})]^{\delta_i} S_t(y_i|\boldsymbol{x}_i, \boldsymbol{\theta})][S_t(w_i|\boldsymbol{x}_i, \boldsymbol{\theta})]^{-1} \times p(\boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i}|\boldsymbol{\alpha})}{\int_{\boldsymbol{x}_{mis,i}} [h_t(y_i|\boldsymbol{x}_i, \boldsymbol{\theta})]^{\delta_i} S_t(y_i|\boldsymbol{x}_i, \boldsymbol{\theta})][S_t(w_i|\boldsymbol{x}_i, \boldsymbol{\theta})]^{-1} \times p(\boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i}|\boldsymbol{\alpha})}. \tag{5.14}$$

which here is a posterior density function. The $(h+1)$th iteration of the EM algorithm would therefore be given by:

1. E-step: compute the expected complete data log-likelihood

$$Q(\boldsymbol{\xi}|\boldsymbol{\xi}^h) = \sum_{i=1}^n E[\ell_i^c(\boldsymbol{\xi}|y_i, \boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i}, \delta_i, w_i)|y_i, \boldsymbol{x}_{obs,i}, \delta_i, w_i] =$$

$$= \sum_{i=1}^n \int_{\boldsymbol{x}_{mis,i}} p(\boldsymbol{x}_{mis,i}|y_i, \boldsymbol{x}_{obs,i}, \delta_i, w_i, \boldsymbol{\xi}^h) \times \ell_i^c(\boldsymbol{\xi}|y_i, \boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i}, \delta_i, w_i, \boldsymbol{\xi}^h), \tag{5.15}$$

2. M-step: find $\boldsymbol{\theta}^{h+1} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^h)$ and set $\boldsymbol{\theta}^{h+1} = \boldsymbol{\theta}^h$,

where $p(\boldsymbol{x}_{mis,i}|y_i, \boldsymbol{x}_{obs,i}, \delta_i, w_i, \boldsymbol{\xi}^h)$ and $\ell_i^c(\boldsymbol{\xi}|y_i, \boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i}, \delta_i, w_i, \boldsymbol{\xi}^h)$ are given by (5.14) and (5.13) respectively, given the current estimates $\boldsymbol{\xi}^h = [\boldsymbol{\theta}^h, \boldsymbol{\alpha}^h]$. If the E-step is not analitically reducible, we may exploit the MCEM algorithm described in Section 4.2.2. In order to implement the MCEM algorithm we have to sample m values for each missing data pattern $\boldsymbol{x}_{mis,i}$. We may draw such samples from the posterior density (5.14) via the Gibbs sampler described in Section 4.3, because of the full conditional assumption (5.9).

It may not be possible to draw samples directly from (5.14), however. In order to bypass such problems we propose the Adaptive Rejection Sampling introduced by Gilks and Wild (1992) and described in Section 4.3.1. The Adaptive Rejection Sampling (ARS) can be applied only if the target density from which we want to sample is log-concave, but allows for discarding the computation of the normalizing constant. In our setting this implies that we can consider the numerator of (5.14) as our target density, which we may

write as

$$
\begin{aligned}
p(\boldsymbol{x}_{mis,i}|y_i, \boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i}, w_i, \boldsymbol{\xi}) \propto\ & p(y_i, \delta_i|\boldsymbol{x}_{obs,i}, \delta_i, w_i, \boldsymbol{\xi}) \\
& \times \quad p(x_{mis,i,k}|y_i, x_{mis,i,k-1}, .., x_{mis,i,r}, \boldsymbol{x}_{obs,i}, \delta_i, w_i, \boldsymbol{\xi}^h) \\
& \times \quad p(x_{mis,i,k-1}|y_i, x_{mis,i,k-2}, .., x_{mis,i,r}, \boldsymbol{x}_{obs,i}, \delta_i, w_i, \boldsymbol{\xi}^h) \\
& \quad \vdots \\
& \times \quad p(x_{mis,i,r+1}|y_i, \boldsymbol{x}_{obs,i}, \delta_i, w_i, \boldsymbol{\xi}^h)
\end{aligned}
$$

assuming that r is the number of observed covariates $\forall i$ and recalling that it is

$$
p(y_i, \delta_i|\boldsymbol{x}_{obs,i}, \delta_i, w_i, \boldsymbol{\xi}) = [h_t(y_i|\boldsymbol{x}_i, \boldsymbol{\theta})]^{\delta_i} S_t(y_i|\boldsymbol{x}_i, \boldsymbol{\theta})][S_t(w_i|\boldsymbol{x}_i, \boldsymbol{\theta})]^{-1} \quad (5.16)
$$

and

$$
\begin{aligned}
p(\boldsymbol{x}_{mis,i}|\boldsymbol{x}_{obs,i}, \boldsymbol{\alpha}) \ =\ & p(x_{mis,i,k}|y_i, x_{mis,i,k-1}, .., x_{mis,i,r}, \boldsymbol{x}_{obs,i}, \delta_i, w_i, \boldsymbol{\xi}^h) \\
& \times \quad p(x_{mis,i,k-1}|y_i, x_{mis,i,k-2}, .., x_{mis,i,r}, \boldsymbol{x}_{obs,i}, \delta_i, w_i, \boldsymbol{\xi}^h) \\
& \quad \vdots \\
& \times \quad p(x_{mis,i,r+1}|y_i, \boldsymbol{x}_{obs,i}, \delta_i, w_i, \boldsymbol{\xi}^h)
\end{aligned}
$$

because of assumption (5.8). Ibrahim, Chen, and MacEachern (1999) proved that (5.16) is log-concave in the components of $\boldsymbol{x}_i$. Recalling that the sum of log-concave densities is still log-concave, we can choose each full conditional density in (5.8) to belong to the exponential family, such that (5.8) is log-concave in the components of $\boldsymbol{x}_i$ as well. Hence, we may sample values for $\boldsymbol{x}_i$ via the Gibbs sampler exploiting the ARS procedure in order to approximate the E-step (5.15). Hence, the whole MCEM algorithm is

1. E-step

   - For each missing data pattern $\boldsymbol{x}_{mis,i}$, draw m samples from (5.14) computed for $\boldsymbol{\theta}^h$ via the Gibbs sampler along with the ARS.

- Compute the approximated expected complete data log-likelihood

$$\widetilde{Q}(\boldsymbol{\xi}|\boldsymbol{\xi}^h) = \frac{1}{m}\sum_{i=1}^{n}\sum_{\boldsymbol{s}_i^j}^{m} \ell_{ij}^c[(\boldsymbol{\xi}|y_i, \boldsymbol{s}_i^j, \boldsymbol{x}_{obs,i}, \delta_i, w_i)|y_i, \boldsymbol{x}_{obs,i}, \delta_i, w_i],$$

  where $\boldsymbol{s}_i^j$ is the set of sampled values for each missing data pattern $\boldsymbol{x}_{mis,i}$, $j = 1, 2, ..., m$.

2. M-step: find $\boldsymbol{\theta}^{h+1} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^h)$, then set $\boldsymbol{\theta}^{h+1} = \boldsymbol{\theta}^h$.

Chapter 6

The proposed model

Random effects are random variables which represent not observable subject-specific determinants. In survival analysis random effects, named frailties, are useful to include unobservable determinants which affect the risk of experiencing an event of interest. Frailties are originally formulated as a multiplicative constant factor on the hazard function which is subject specific . Such a model is also called over-dispersion model. To specify the model correctly, simply the marginal distribution of failure times is needed, which is obtained integrating out the frailty (see Section 3.2).

The original frailty model has been developed in different directions. For example Hougaard (1984, 1986) proposed several new possible distributions for the frailty discussing their properties and the consequences of their use. Petersen (1998) proposed a frailty model where the impact on hazard is given by the sum of two or more gamma frailty terms. Dynamic frailties were presented in Yue and Chan (1997) where individual random effects' values are allowed to vary over time stochastically. In Section 3.3 we described the proportional hazards model with random effects, which allows for group-specific random effects.

Despite the great variety of such random effects' models there is a common element which has been unchanged: the random effects are always considered something which is totally independent of the covariates. This assumption

may be too restrictive sometimes. For example,the risk for a child to become asthmatic is due both to a genetic inheritance and external factors, i.e. the mother is a smoker. Usually this is modeled via a frailty component and a predictor, both multiplied to a baseline risk. This implies that the genetic and the external factors are independent. However, it may be that the smoking of mother during pregnancy affected even the genetic component, e.g if the mother used to smoke before and during pregnancy. If this is the case, a traditional frailty model would not capture this effect, even if the predictor would be correctly specified including in it the covariate which denotes the mother smoking status. The effect of smoke on the genetic component could be modeled assuming that the expected value of the child frailty distribution is a function of the covariate which denotes the mother smoking status. Another example can be made in economics. Let us assume we want to study the time to first job after the master degree. The risk of experiencing the event at time t could be affected by several observable factors as sex, age, type of degree, so we include these covariates in the model. Beside them, subject-specific unobservable determinants may affect the hazard as well, like individuals' own determination in looking for a job. The influence of these factors may be modeled by an over-dispersion frailty model, where a subject-specific frailty multiplies the hazard function. It seems reasonable to assume that individual determination may depend on sex, that is women could be more resoluted than men in looking for a job because they expect to be discriminated. In more general terms, it may be that subject-specific unobservable determinants are someway affected by one or more observable features. In order to include this relation into a frailty model, we propose a model which introduces a dependence of the subject-specific frailty distribution on one or more observable covariates. This assumption is general with respect to the specification of hazard. In order to estimate our model we exploit the estimation procedure for PH model with missing values in covariates described in Chapter 5.

6.1 The traditional framework

Let T be a positive random variable (RV) from which the failure times $\{t_i\}$ of n subjects are independently drawn, $i = 1, 2, ..., n$. Let also define the positive RVs C and W, from which respectively censoring times c_i and entering times w_i are independently drawn. Because of possible left-truncation and right-censoring we observe the event time $Y = \min(T, C)$ such that a faliure indicator δ_i equals 1, if a failure occurs, 0 otherwise. Let each individual be described by a vector of fully observed covariates $\boldsymbol{x}_i = [x_{i1}, ..., x_{ik}]'$ and $\boldsymbol{\beta}$ a vector of parameters.

Now let us assume that the n observed subjects are *heterogeneus* with respect to some unobserved subject-specific feature which impacts on failure times $\{t_i\}$. Let such unobserved heterogeneity be included in the model via n univariate subject-specific random effects $\{f_i\}$ drawn from a positive univariate RV $F \sim (\cdot|\boldsymbol{\varsigma})$. Recalling the proportional hazards model with random effects described in Section 3.3, we may write the individual hazard function (3.6) as

$$h(t|\boldsymbol{x}_i, \boldsymbol{\beta}, f_i) = h_0(t) \exp(\boldsymbol{\beta}'\boldsymbol{x}_i + f_i)$$

which is given by (3.6) setting $P = n$, $p = i$, $q = 1$ and $\omega_{p,i} = 1 \, \forall i$. Allowing for right-censoring the risk of subject i to have an event at time y_i may be written as

$$h(y_i|\boldsymbol{x}_i, f_i, \boldsymbol{\theta}) = h_0(y_i|\boldsymbol{\eta}) \cdot \exp(\boldsymbol{\beta}'\boldsymbol{x}_i + f_i), \qquad \boldsymbol{\theta} = [\boldsymbol{\eta}, \boldsymbol{\beta}].$$

and the ith likelihood contribution as

$$L_i(\boldsymbol{\theta}, \boldsymbol{\varsigma}|y_i, \boldsymbol{x}_i, f_i, \delta_i, w_i) =$$

$$= \int_{f_i} \Big\{ [h_t(y_i|\boldsymbol{x}_i, f_i, \boldsymbol{\theta})]^{\delta_i} S_t(y_i|\boldsymbol{x}_i, f_i, \boldsymbol{\theta}) \times [S_t(w_i|\boldsymbol{x}_i, f_i, \boldsymbol{\theta})]^{-1} \Big\} dp(f_i|\boldsymbol{\varsigma}).$$

6.2 The random effects' distribution

In the previous section we assumed the random effects $\{f_i\}$ to be n i.i.d. realizations of the same random variable F, that is we assumed $p(\cdot|\boldsymbol{\varsigma})$ to be the *same* probability distribution function from which the individual random effects were drawn. It would be equivalent to assume that the n random effects are independently drawn from n identical prior probability distribution functions, i.e $p(\cdot|\boldsymbol{\varsigma}) = p(\cdot|\boldsymbol{\varsigma}_i)$ assuming $\boldsymbol{\varsigma}_i = \boldsymbol{\varsigma} \quad \forall i$.

Now let us assume

$$\boldsymbol{\varsigma}_i \neq \boldsymbol{\varsigma}_{i'}, \qquad \forall i \neq i',\ i, i' = 1, ..., n. \tag{6.1}$$

that is each $p(\cdot|\boldsymbol{\varsigma}_i)$ has the *same* functional form but distinct parameters' vectors. Let us introduce the further assumption

$$\boldsymbol{\varsigma}_i = g(\tilde{\boldsymbol{x}}_i, \boldsymbol{\alpha}), \tag{6.2}$$

where $g(\cdot)$ is an arbitrary function. The vector of parameters $\boldsymbol{\alpha} \neq \boldsymbol{\theta}$ is the same for each f_i, while $\boldsymbol{x}_i'$ denotes a vector of covariates which are choosen among the k covariates of subject i. Assumptions (6.1) and (6.2) imply that $p(f_i|\boldsymbol{\varsigma}_i)$ depends on the ith set of choosen covariates via $\boldsymbol{\alpha}$, so we may write

$$f_i \sim p(f_i|\tilde{\boldsymbol{x}}_i, \boldsymbol{\alpha}). \tag{6.3}$$

6.3 Estimation of the model

The random effect f_i may be considered as a missing value $\forall i$ if we look at the structure of the data. Thus, we may think about applying the EM estimation procedure for PH model with missing values described in Chapter 5 in order to estimate the parameters of interest.

We may refer to $\boldsymbol{x}_i$ as the incomplete data set for the ith subject, while $[\boldsymbol{x}_i, f_i]'$ is the complete data set. Recalling the likelihood (5.2) we may write

the ith incomplete data likelihood contribution as

$$
\begin{aligned}
&L_i(\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\phi}, \boldsymbol{\alpha} | y_i, \boldsymbol{x}_i, f_i, \delta_i, w_i, \boldsymbol{r}_i) = \\
&= \int_{f_i} \Big\{ p_r(\boldsymbol{r}_i | y_i, \boldsymbol{x}_i, f_i, \delta_i, w_i, \boldsymbol{\phi}) \times [h_t(y_i | \boldsymbol{x}_i, f_i, \boldsymbol{\theta})]^{\delta_i} S_t(y_i | \boldsymbol{x}_i, f_i, \boldsymbol{\theta}) \\
&\quad \times [S_t(w_i | \boldsymbol{x}_i, f_i, \boldsymbol{\theta})]^{-1} [h_c(y_i | \boldsymbol{x}_i, f_i, \boldsymbol{\psi})]^{1-\delta_i} S_c(y_i | \boldsymbol{x}_i, f_i, \boldsymbol{\psi}) \Big\} dp(f_i | \tilde{\boldsymbol{x}}_i, \boldsymbol{\alpha}).
\end{aligned} \tag{6.4}
$$

We recall that $h_c(\cdot|\boldsymbol{x}_i, f_i, \boldsymbol{\psi})$ and $S_c(\cdot|\boldsymbol{x}_i, f_i, \boldsymbol{\psi})$ are the subject-specific conditional hazard and survival function of censoring time, while $p(\cdot|\boldsymbol{\varsigma})$ is the pdf of $(\boldsymbol{x}_i, f_i)$. Furthermore, $\boldsymbol{r}_i$ is a $(k+1)$-vector in which the component r_{ij} is 1 if x_{ij}, f_i is observed, 0 otherwise, $j = 0, 1, .., k$. The vector $\boldsymbol{r}_i$ denotes the ith individual missingness profile with pdf $p_r(\boldsymbol{r}_i | y_i, \boldsymbol{x}_i, f_i, \delta_i, w_i, \boldsymbol{\phi})$, $\boldsymbol{\phi}$ a vector of parameters. In the current setting, it is $\boldsymbol{r}_i = [1, ..., 1, 0]' \quad \forall i$, where the 1s and the 0 correspond to $x_1, ..., x_k$ and f_i respectively. Note that the probability that we do not observe the random effect f_i is independent of $\boldsymbol{x}_i$, i.e. f_i is *missing completely at random* (see Chapter 5). This implies that

$$
p_r(\boldsymbol{r}_i | y_i, \boldsymbol{x}_i, f_i, \delta_i, w_i, \boldsymbol{\phi}) = p_r(\boldsymbol{r}_i | y_i, \delta_i, w_i, \boldsymbol{\phi}). \tag{6.5}
$$

Let us assume that the censoring distribution is independent of missing values, that is

$$
h_c(y_i | \boldsymbol{x}_i, f_i, \boldsymbol{\psi})^{1-\delta_i} S_c(y_i | \boldsymbol{x}_i, f_i, \boldsymbol{\psi}) = h_c(y_i | \boldsymbol{x}_i, \boldsymbol{\psi})^{1-\delta_i} S_c(y_i | \boldsymbol{x}_i, \boldsymbol{\psi}). \tag{6.6}
$$

as well as $\boldsymbol{\theta}, \boldsymbol{\varsigma} \neq \boldsymbol{\phi}, \boldsymbol{\psi}$. The contribution (6.4) becomes

$$
\begin{aligned}
&L_i(\boldsymbol{\theta}, \boldsymbol{\alpha} | y_i, \boldsymbol{x}_i, f_i, \delta_i, w_i) = \\
&= \int_{f_i} \Big\{ [h_t(y_i | \boldsymbol{x}_i, f_i, \boldsymbol{\theta})]^{\delta_i} S_t(y_i | \boldsymbol{x}_i, f_i, \boldsymbol{\theta}) \times [S_t(w_i | \boldsymbol{x}_i, f_i, \boldsymbol{\theta})]^{-1} \Big\} dp(f_i | \tilde{\boldsymbol{x}}_i, \boldsymbol{\alpha}).
\end{aligned} \tag{6.7}
$$

as (6.5) and (6.6) are constant with respect to the integral (6.4).

Thus, in order to estimate $\boldsymbol{\xi}$ we may maximize the expected complete data log-likelihood via the EM algorithm. The expectation is taken with respect to the conditional distribution of each frailty component f_i given the observed data, which consists of $(y_i, \tilde{\boldsymbol{x}}_i, \delta_i, w_i)$ for each individual. In the remainder of this section we describe the EM procedure for a discrete and a continuous random effect respectively.

6.3.1 Discrete random effect

If we assume that f_i is discrete $\forall i$, the ith expected complete data log-likelihood is given by

$$E[\ell_i^c(\boldsymbol{\xi}|y_i, \boldsymbol{x}_i, f_i, \delta_i, w_i)|y_i, \boldsymbol{x}_i, \delta_i, w_i] =$$

$$= \sum_{f_i} p(f_i|y_i, \boldsymbol{x}_i, \delta_i, w_i, \boldsymbol{\xi}) \times \ell_i^c(\boldsymbol{\xi}|y_i, \boldsymbol{x}_i, f_i, \delta_i, w_i)$$

where

$$\ell_i^c(\boldsymbol{\xi}|y_i, \boldsymbol{x}_i, f_i, \delta_i, w_i) =$$

$$= \log\left\{[h_t(y_i|\boldsymbol{x}_i, f_i, \boldsymbol{\theta})]^{\delta_i} S_t(y_i|\boldsymbol{x}_i, f_i, \boldsymbol{\theta})[S_t(w_i|\boldsymbol{x}_i, f_i, \boldsymbol{\theta})]^{-1} \times p(f_i|\tilde{\boldsymbol{x}}_i, \boldsymbol{\alpha})\right\} \tag{6.8}$$

and

$$p(f_i|y_i, \boldsymbol{x}_i, \delta_i, w_i, \boldsymbol{\xi}) =$$

$$= \frac{[h_t(y_i|\boldsymbol{x}_i, f_i, \boldsymbol{\theta})]^{\delta_i} S_t(y_i|\boldsymbol{x}_i, f_i, \boldsymbol{\theta})[S_t(w_i|\boldsymbol{x}_i, f_i, \boldsymbol{\theta})]^{-1} \times p(f_i|\tilde{\boldsymbol{x}}_i, \boldsymbol{\alpha})}{\sum_{f_i}[h_t(y_i|\boldsymbol{x}_i, f_i, \boldsymbol{\theta})]^{\delta_i} S_t(y_i|\boldsymbol{x}_i, f_i, \boldsymbol{\theta})[S_t(w_i|\boldsymbol{x}_i, f_i, \boldsymbol{\theta})]^{-1} \times p(f_i|\tilde{\boldsymbol{x}}_i, \boldsymbol{\alpha})}. \tag{6.9}$$

The score equations to solve are

$$\boldsymbol{u}^*(\boldsymbol{\xi}) = \sum_i \sum_{f_i} p_{f_i} \times \begin{bmatrix} \boldsymbol{u}_{\boldsymbol{\beta},i}(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} \ell_i^c(\boldsymbol{\xi}|y_i, \boldsymbol{x}_i, f_i, \delta_i, w_i) \\ \boldsymbol{u}_{\boldsymbol{\eta},i}(\boldsymbol{\eta}) = \frac{\partial}{\partial \boldsymbol{\eta}} \ell_i^c(\boldsymbol{\xi}|y_i, \boldsymbol{x}_i, f_i, \delta_i, w_i) \\ \boldsymbol{u}_{\boldsymbol{\alpha},i}(\boldsymbol{\alpha}) = \frac{\partial}{\partial \boldsymbol{\alpha}} \ell_i^c(\boldsymbol{\xi}|y_i, \boldsymbol{x}_i, f_i, \delta_i, w_i) \end{bmatrix} = \mathbf{0},$$

where p_{f_i} is given by (6.9)

After having initialized the parameters, at the $(h+1)$th iteration the steps of the EM algorithm are:

1. E-step

   - **for** each subject i
   - **for** each possible value of f_i
   - compute the expected complete data log-likelihood contribution of subject i

     $$\sum_{f_i} p(f_i|y_i, \boldsymbol{x}_i, \delta_i, w_i, \boldsymbol{\xi}^h) \times \ell_i^c(\boldsymbol{\xi}|y_i, \boldsymbol{x}_i, f_i, \delta_i, w_i, \boldsymbol{\xi}^h).$$

   - compute the expected complete data log-likelihood

     $$Q(\boldsymbol{\xi}|\boldsymbol{\xi}^h) = \sum_i \sum_{f_i} p(f_i|y_i, \boldsymbol{x}_i, \delta_i, w_i, \boldsymbol{\xi}^h) \times \ell_i^c(\boldsymbol{\xi}|y_i, \boldsymbol{x}_i, f_i, \delta_i, w_i, \boldsymbol{\xi}^h)$$

2. Estimate $\boldsymbol{\xi}^{h+1}$ by maximizing $Q(\boldsymbol{\xi}|\boldsymbol{\xi}^h)$ with respect to $\boldsymbol{\xi}$,

$$\boldsymbol{\xi}^{h+1} = \arg\max_{\boldsymbol{\xi}} Q(\boldsymbol{\xi}|\boldsymbol{\xi}^h)$$

- Update the set of parameters setting $\boldsymbol{\xi}^h = \boldsymbol{\xi}^{h+1}$

Let $\hat{\boldsymbol{\xi}} = [\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}}]'$ be the estimates of $\boldsymbol{\xi} = [\boldsymbol{\alpha}, \boldsymbol{\theta}]'$ after the EM algorithm is completed, where $\boldsymbol{\theta} = [\boldsymbol{\eta}, \boldsymbol{\beta}]'$ is the vector of the parameters of interest. Along with Herring and Ibrahim (2001), we propose the robust "sandwich" estimator in order to estimate the variance of $\hat{\boldsymbol{\theta}}$ (Pugh et al., 1993):

$$J_n(\hat{\boldsymbol{\theta}})^{-1} V_n(\hat{\boldsymbol{\theta}}) J_n(\hat{\boldsymbol{\theta}})^{-1}, \tag{6.10}$$

where

$$\begin{aligned} V_n(\hat{\boldsymbol{\theta}}) &= \frac{1}{n}\Bigg(\sum_{i=1}^{n}\sum_{j=1}^{m} \hat{p}_{f_{ij}} \boldsymbol{u}_{ij}(\hat{\boldsymbol{\theta}}) \boldsymbol{u}'_{ij}(\hat{\boldsymbol{\theta}}) - \left[\sum_{i=1}^{n}\sum_{j=1}^{m} \hat{p}_{f_{ij}} \boldsymbol{u}_{ij}(\hat{\boldsymbol{\theta}}) \boldsymbol{u}'_{ij}(\hat{\boldsymbol{\alpha}}) \right] \\ &\times \left[\sum_{i=1}^{n}\sum_{j=1}^{m} \hat{p}_{f_{ij}} \boldsymbol{u}_{ij}(\hat{\boldsymbol{\alpha}}) \boldsymbol{u}'_{ij}(\hat{\boldsymbol{\alpha}}) \right]^{-1} \times \left[\sum_{i=1}^{n}\sum_{j=1}^{m} \hat{p}_{f_{ij}} \boldsymbol{u}_{ij}(\hat{\boldsymbol{\alpha}}) \boldsymbol{u}'_{ij}(\hat{\boldsymbol{\theta}}) \right] \Bigg) \end{aligned} \tag{6.11}$$

and

$$J_n(\hat{\boldsymbol{\theta}}) = \frac{1}{n} \sum_{i=1}^{n}\sum_{j=1}^{m} \hat{p}_{f_{ij}} \nabla^2 \ell^c_{ij}(\hat{\boldsymbol{\xi}}), \qquad \hat{\boldsymbol{\theta}} = [\hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\beta}}]'. \tag{6.12}$$

In $V_n(\hat{\boldsymbol{\theta}})$, j indexes the possible values that f_i can assume $\forall i$. Furthermore, $\hat{p}_{f_{ij}}$ is the posterior probability given by (6.9) computed in $\hat{\boldsymbol{\xi}} = [\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\alpha}}]'$. The vector $\boldsymbol{u}_{ij}(\hat{\boldsymbol{\theta}}) = [\boldsymbol{u}_{\boldsymbol{\beta},ij}(\hat{\boldsymbol{\beta}}), \boldsymbol{u}_{\boldsymbol{\eta},ij}(\hat{\boldsymbol{\eta}})]'$ is the vector of the ith contributions to the score functions with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$, computed in $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\eta}}$ and for each possible value of f_i. Similarly, $\boldsymbol{u}_{ij}(\hat{\boldsymbol{\alpha}})$ is given by the contribution to the score function's vector $\boldsymbol{u}_{\boldsymbol{\alpha},i}$ computed in $\hat{\boldsymbol{\alpha}}$ for each possible value of f_i. The quantity $\ell^c_{ij}(\hat{\boldsymbol{\xi}})$ is the ith complete data log-likelihood contribution (6.8)

computed for each value of f_i indexed by j.

6.3.2 Continuous random effect

If the frailty f_i is continuous $\forall i$ the individual expected complete data contribution is given by

$$E[\ell_i(\boldsymbol{\xi}|y_i, \boldsymbol{x}_i, f_i, \delta_i, w_i)|y_i, \boldsymbol{x}_i, \delta_i, w_i] =$$

$$\int_{f_i} p(f_i|y_i, \boldsymbol{x}_i, \delta_i, w_i, \boldsymbol{\xi}) \times \ell_i^c(\boldsymbol{\xi}|y_i, \boldsymbol{x}_i, f_i, \delta_i, w_i) df_i \tag{6.13}$$

where

$$\ell_i^c(\boldsymbol{\xi}|y_i, \boldsymbol{x}_i, f_i, \delta_i, w_i) =$$

$$= \log\left\{[h_t(y_i|\boldsymbol{x}_i, f_i, \boldsymbol{\theta})]^{\delta_i} S_t(y_i|\boldsymbol{x}_i, f_i, \boldsymbol{\theta})[S_t(w_i|\boldsymbol{x}_i, f_i, \boldsymbol{\theta})]^{-1} \times p(f_i|\tilde{\boldsymbol{x}}_i, \boldsymbol{\alpha})\right\}.$$

is the complete data log-likelihood contribution and $p(f_i|\tilde{\boldsymbol{x}}_i, \boldsymbol{\alpha})$ is the density function of f_i. Furthermore,

$$p(f_i|y_i, \boldsymbol{x}_i, \delta_i, w_i, \boldsymbol{\xi}) =$$

$$= \frac{[h_t(y_i|\boldsymbol{x}_i, f_i, \boldsymbol{\theta})]^{\delta_i} S_t(y_i|\boldsymbol{x}_i, f_i, \boldsymbol{\theta})[S_t(w_i|\boldsymbol{x}_i, f_i, \boldsymbol{\theta})]^{-1} \times p(f_i|\tilde{\boldsymbol{x}}_i, \boldsymbol{\alpha})}{\int_{f_i}[h_t(y_i|\boldsymbol{x}_i, f_i, \boldsymbol{\theta})]^{\delta_i} S_t(y_i|\boldsymbol{x}_i, f_i, \boldsymbol{\theta})[S_t(w_i|\boldsymbol{x}_i, f_i, \boldsymbol{\theta})]^{-1} \times p(f_i|\tilde{\boldsymbol{x}}_i, \boldsymbol{\alpha}) df_i}. \tag{6.14}$$

is the posterior conditional density of f_i. The score equations are

$$\boldsymbol{u}^*(\boldsymbol{\xi}) = \sum_{i=1}^{n} \int_{f_i} p_{f_i} \times \begin{bmatrix} \boldsymbol{u}_{\boldsymbol{\beta},i}(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} \ell_i^c(\boldsymbol{\xi}|y_i, \boldsymbol{x}_i, f_i, \delta_i, w_i) \\ \boldsymbol{u}_{\boldsymbol{\eta},i}(\boldsymbol{\eta}) = \frac{\partial}{\partial \boldsymbol{\eta}} \ell_i^c(\boldsymbol{\xi}|y_i, \boldsymbol{x}_i, f_i, \delta_i, w_i) \\ \boldsymbol{u}_{\boldsymbol{\alpha},i}(\boldsymbol{\alpha}) = \frac{\partial}{\partial \boldsymbol{\alpha}} \ell_i^c(\boldsymbol{\xi}|y_i, \boldsymbol{x}_i, f_i, \delta_i, w_i) \end{bmatrix} df_i = \mathbf{0}. \tag{6.15}$$

where p_{f_i} is given by (6.14).

If the integral in (6.13) does not have a closed form we may exploit the MCEM algorithm described in Section 4.2.2, that is we may approximate (6.13) by

$$\widetilde{E}[\ell_i^c(\boldsymbol{\xi}|y_i, \boldsymbol{x}_i, f_i, \delta_i, w_i)|y_i, \boldsymbol{x}_i, \delta_i, w_i] = \frac{1}{m} \sum_{r=q+1}^{R} \ell_i^c(\boldsymbol{\xi}|y_i, \boldsymbol{x}_i, s_{i,r}, \delta_i, w_i).$$

The quantity $s_{i,r}^h$ is the rth element of the vector $\boldsymbol{s}_i^h$ of sampled values for f_i and q is the number of initial samples which are discarded (burn-in). The samples may be drawn from (6.14). Recalling what we wrote in Section 5.3, if we choose $p(f_i|\widetilde{\boldsymbol{x}}_i, \boldsymbol{\alpha})$ to belong to the exponential family we may exploit the Adaptive Rejection Sampling algorithm to draw samples from (6.14).

In this case we have only one conditional density to sample from, which derives from the fact that we do not need to specify the full conditionals for the observed covariates.

After having set a vector of initial values $\boldsymbol{\xi}^0$, at the $(h+1)$th iteration the steps of the MCEM algorithm are

1. E-step:

   - **for** each subject i

- Sample $m = p - q$ values for f_i from

$$p(f_i|y_i, \boldsymbol{x}_i, \delta_i, w_i, \boldsymbol{\xi}^h)$$

where p is the total number of sampled values.

- compute the approximated expected complete data log-likelihood contribution

$$\widetilde{E}[\ell_i(\boldsymbol{\xi}|y_i, \boldsymbol{x}_i, f_i, \delta_i, w_i)|y_i, \boldsymbol{x}_i, \delta_i, w_i] = \frac{1}{m}\sum_{r=q+1}^{R} \ell_i^c(\boldsymbol{\xi}|y_i, \boldsymbol{x}_i, s_{i,r}^h, \delta_i, w_i),$$

- compute the expected complete data log-likelihood

$$Q(\boldsymbol{\xi}|\boldsymbol{\xi}^h) = \frac{1}{m}\sum_i \sum_{r=q+1}^{R} \ell_i^c(\boldsymbol{\xi}|y_i, \boldsymbol{x}_i, s_{i,r}^h, \delta_i, w_i, \boldsymbol{\xi}^h)$$

2. M-step: estimate $\boldsymbol{\xi}^{h+1}$ by maximizing $Q(\boldsymbol{\xi}|\boldsymbol{\xi}^h)$ with respect to $\boldsymbol{\xi}$

$$\boldsymbol{\xi}^{h+1} = \arg\max_{\boldsymbol{\xi}} Q(\boldsymbol{\xi}|\boldsymbol{\xi}^h)$$

- Update the set of parameters setting $\boldsymbol{\xi}^h = \boldsymbol{\xi}^{h+1}$.

As suggested by Herring et al. (2002), when the random effect is continuous we may use the variance estimator for $\boldsymbol{\theta}$ proposed by Goetgheburger and Ryan (2000) following Rubin and Schenker (1991). Such estimator is simple and presents good small sample properties. Let $\boldsymbol{\theta} = [\boldsymbol{\eta}, \boldsymbol{\beta}]'$ be the vector of our parameters of interest and $\hat{\boldsymbol{\theta}}$ the estimates of $\boldsymbol{\theta}$ after the MCEM algorithm is completed. Let also $\hat{\boldsymbol{\xi}} = [\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}}]'$ be the estimates of $\boldsymbol{\xi} = [\boldsymbol{\alpha}, \boldsymbol{\theta}]'$ after the MCEM algorithm is completed. Then, for each ith subject we draw m values of f_i from $p(f_i|y_i, \boldsymbol{x}_i, \delta_i, \omega_i, \hat{\boldsymbol{\xi}})$. We define the $n \times m$ matrix $\boldsymbol{S}$ where

each jth column is an n-vector $\boldsymbol{s}_j$ of imputed values for the missing covariate f, $j = 1, ...m$. Then we estimate m different vectors of parameters $\hat{\boldsymbol{\theta}}_j$ by maximizing m different log-likelihoods

$$\ell_j(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{s}_j, \boldsymbol{\delta}, \boldsymbol{\omega}) = \sum_{i=1}^{n} \ell_{i,j}(\boldsymbol{\theta}|y_i, \boldsymbol{x}_i, s_{i,j}, \delta_i, w_i)$$

with respect to $\boldsymbol{\theta}$, where $s_{i,j}$ is the jth sampled value of f_i. For each $\hat{\boldsymbol{\theta}}_j$ we calculate the *within*-imputation variance $W_j = I_j^{-1}(\hat{\boldsymbol{\theta}})$ where $I_j(\hat{\boldsymbol{\theta}})$ is the observed information matrix associated to each $\hat{\boldsymbol{\theta}}_j$. We compute then the mean of variances $(W_1, ..., W_j, ..., W_m)$,

$$\bar{W} = \sum_{j=1}^{m} \frac{W_j}{m}.$$

Let also

$$\bar{\boldsymbol{\theta}}_j = \sum_{j=1}^{m} \frac{\hat{\boldsymbol{\theta}}_j}{m}$$

be the average of the m estimates $\hat{\boldsymbol{\theta}}_j$. Then we calculate the *between*-imputation variance

$$B = \frac{(\hat{\boldsymbol{\theta}}_j - \bar{\boldsymbol{\theta}}_j)'(\hat{\boldsymbol{\theta}}_j - \bar{\boldsymbol{\theta}}_j)}{m-1}.$$

The variance estimator is then given by

$$\hat{V} = \bar{W} + (1 + m^{-1})B. \tag{6.16}$$

Chapter 7

Simulations

In this chapter we describe the results of two different simulations sets we performed, according to the proposed model described in Chapter 6. The first set of simulations was performed assuming a binary random effect. In the second set of simulations we assumed a continuous random effect. We estimated the models along with the EM and MCEM procedures described in Sections 6.3.1 and 6.3.2 respectively. In each of both cases we compared the results with two different models. The first one is a model which does not account for the presence of the random effect, which we named *without frailty* model (WF). The second model is the *full data* model (FD), which is the model in which we assume to know the values of the random effect. We replicated 100 independent simulations both for the binary and the continuous RE to provide coverage probabilities for the estimates.

7.1 Binary random effect

We built up a set of $n = 1000$ subjects each described by two covariates x_{i1} and x_{i2} which are observed $\forall i$ and are independently drawn from a random

variable $X \sim N(0,1)$. We assumed a Weibull baseline hazard,

$$h_0(t|\boldsymbol{\eta}) = \gamma\lambda t^{\gamma-1}, \qquad \boldsymbol{\eta} = [\gamma, \lambda] \tag{7.1}$$

where $\gamma = 0.6$ and $\lambda = 0.3$. Each subject was assigned a random effect f_i. According to the assumptions made in Section 6.2 each random effect f_i was independently drawn from a subject-specific distribution. In this setting it was

$$f_i \sim Bernoulli(p_i),$$

where we assumed

$$p_i = \frac{\exp(\alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2})}{1 + \exp(\alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2})} \tag{7.2}$$

and $\alpha_0 = 0.2$, $\alpha_1 = 0.3$, $\alpha_2 = 0.6$. We generated n event times $\{t_i\}$ randomly, according to the subject-specific conditional hazard

$$h(y_i, \delta_i|\boldsymbol{x}_i, f_i, \boldsymbol{\theta}) = \gamma\lambda y_i^{\gamma-1} \exp(\beta_1 x_{i1} + \beta_2 x_{i2} + f_i), \qquad \boldsymbol{\theta} = [\boldsymbol{\eta}, \boldsymbol{\beta}]'$$

setting $\boldsymbol{\beta} = [\beta_1, \beta_2]' = [0.7, 0.5]'$. Namely it was

$$t_i = \left\{-[\lambda \exp(\beta_1 x_{i1} + \beta_2 x_{i2} + f_i)] \log[1-u]\right\}^{-\frac{1}{\gamma}},$$

where $u \sim U(0,1)$ is the random value of the cumulative distribution function of the lifetime t_i given by

$$F(t_i|\boldsymbol{x}_i, f_i, \boldsymbol{\theta}) = 1 - S(t_i|\boldsymbol{x}_i, f_i, \boldsymbol{\theta}) = 1 - \exp[-\lambda t_i^{\gamma} \exp(\beta_1 x_{i1} + \beta_2 x_{i2} + f_i)].$$

We allowed for both right-censoring and left-truncation by drawing the censoring times $\{c_i\}$ and the entry times w_i independently from two independent RVs C and W.

We defined the misspecified model without considering the presence of

the random effect f_i. Under such model the ith log-likelihood contribution was

$$\begin{aligned}\ell_i(\boldsymbol{\theta}|y_i, \boldsymbol{x}_i, \delta_i, w_i) &= \delta_i[\log(\gamma\lambda y_i^{\gamma-1}) + \beta_1 x_{i1} + \beta_2 x_{i2}] \\ &+ \lambda \exp(\beta_1 x_{i1} + \beta_2 x_{i2})(w_i^\gamma - y_i^\gamma), \qquad \boldsymbol{\theta} = [\gamma, \lambda, \beta_1, \beta_2]',\end{aligned} \tag{7.3}$$

where $y_i = \min(t_i, c_i)$ is the observed exit time and δ_i is the indicator which equals 1 if $y_i = t_i$, 0 otherwise. Assuming we knew the real value of $f_i \, \forall i$ we defined thr ith log-likelihood contribution of the full data model (FD) as

$$\begin{aligned}\ell_i(\boldsymbol{\theta}|y_i, \boldsymbol{x}_i, \delta_i, w_i) &= \delta_i[\log(\gamma\lambda y_i^{\gamma-1}) + \beta_1 x_{i1} + \beta_2 x_{i2} + f_i] \\ &+ \lambda \exp(\beta_1 x_{i1} + \beta_2 x_{i2} + f_i)(w_i^\gamma - y_i^\gamma).\end{aligned}$$

In both cases we achieve the MLEs $\hat{\boldsymbol{\theta}}$ by solving the score equations

$$\boldsymbol{u}(\boldsymbol{\theta}) = \sum_i^n \begin{bmatrix} u_{\gamma,i}(\gamma) = \frac{\partial}{\partial\gamma}\ell_i(\boldsymbol{\theta}|y_i, \boldsymbol{x}_i, \delta_i, w_i) \\ u_{\lambda,i}(\lambda) = \frac{\partial}{\partial\lambda}\ell_i(\boldsymbol{\theta}|y_i, \boldsymbol{x}_i, \delta_i, w_i) \\ u_{\beta_1,i}(\beta_1) = \frac{\partial}{\partial\beta_1}\ell_i(\boldsymbol{\theta}|y_i, \boldsymbol{x}_i, \delta_i, w_i) \\ u_{\beta_2,i}(\beta_2) = \frac{\partial}{\partial\beta_2}\ell_i(\boldsymbol{\theta}|y_i, \boldsymbol{x}_i, \delta_i, w_i) \end{bmatrix} = \mathbf{0}. \tag{7.4}$$

Under the WF model the score functions in (7.4) are given by

$$\begin{aligned}
u_\gamma(\gamma) &= \sum_{i=1}^{n}\{\delta_i\left[\gamma^{-1}+\log(y_i)\right]+\exp(\beta_1 x_{1i}+\beta_2 x_{2i}) \\
&\qquad\qquad \times\lambda\left[w_i^\gamma\log(w_i^\gamma)-y_i^\gamma\log(y_i^\gamma)\right]\}, \\
u_\lambda(\lambda) &= \sum_{i=1}^{n}[\delta_i\lambda^{-1}+\exp(\beta_1 x_{1i}+\beta_2 x_{2i})\left(w_i^\gamma-y_i^\gamma\right)], \\
u_{\beta_1}(\beta_1) &= \sum_{i=1}^{n}[\delta_i x_{1i}+\lambda(w_i^\gamma-y_i^\gamma)\exp(\beta_1 x_{1i}+\beta_2 x_{2i})x_{1i}], \\
u_{\beta_2}(\beta_2) &= \sum_{i=1}^{n}[\delta_i x_{2i}+\lambda(w_i^\gamma-y_i^\gamma)\exp(\beta_1 x_{2i}+\beta_2 x_{2i})x_{2i}].
\end{aligned}$$

The score functions under the FD model are given by (7.9)-(7.12) at page 77.

Now we consider the EM estimation procedure to account for unobserved heterogeneity. We recall that we can not observe f_i so we need to consider the prior distribution of f_i to specify likelihood correctly. In the current setting it was

$$\begin{aligned}
&p(f_i|\boldsymbol{x}_i,\boldsymbol{\alpha}) = \\
&= \left[\frac{\exp(\alpha_0+\alpha_1 x_{i1}+\alpha_2 x_{i2})}{1+\exp(\alpha_0+\alpha_1 x_{i1}+\alpha_2 x_{i2})}\right]^{f_i}\left[1-\frac{\exp(\alpha_0+\alpha_1 x_{i1}+\alpha_2 x_{i2})}{1+\exp(\alpha_0+\alpha_1 x_{i1}+\alpha_2 x_{i2})}\right]^{1-f_i}.
\end{aligned}$$

As we wrote in Section 6.3.1 the ith log-likelihood contribution may be written as the expected complete data log-likelihood contribution

$$\begin{aligned}
&E[\ell_i^c(\boldsymbol{\xi}|y_i,\boldsymbol{x}_i,f_i,\delta_i,w_i)|y_i,\boldsymbol{x}_i,\delta_i,w_i] = \\
&\qquad = \sum_{f_i} p(f_i|y_i,\boldsymbol{x}_i,\delta_i,w_i,\boldsymbol{\xi})\times\ell_i^c(\boldsymbol{\xi}|y_i,\boldsymbol{x}_i,f_i,\delta_i,w_i), \qquad (7.5)
\end{aligned}$$

where

$$
\begin{aligned}
& p(f_i|y_i, \boldsymbol{x}_i, \delta_i, w_i, \boldsymbol{\xi}) = \\
& \quad \gamma\lambda y_i^{\gamma-1} \exp(\beta_1 x_{i1} + \beta_2 x_{i2} + f_i)]^{\delta_i} \exp[\lambda(w_i^\gamma - y_i^\gamma) \exp(\beta_1 x_{i1} + \beta_2 x_{i2} + f_i)] \\
\times & \quad \left[\frac{\exp(\alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2})}{1 + \exp(\alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2})}\right]^{f_i} \left[1 - \frac{\exp(\alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2})}{1 + \exp(\alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2})}\right]^{1-f_i} \\
\times & \quad \left\{ \sum_{f_i} y_i^{\gamma-1} \exp(\beta_1 x_{i1} + \beta_2 x_{i2} + f_i)]^{\delta_i} \exp[\lambda(w_i^\gamma - y_i^\gamma) \exp(\beta_1 x_{i1} + \beta_2 x_{i2} + f_i)] \right. \\
& \quad \left. \times \left[\frac{\exp(\alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2})}{1 + \exp(\alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2})}\right]^{f_i} \left[1 - \frac{\exp(\alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2})}{1 + \exp(\alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2})}\right]^{1-f_i} \right\}^{-1}
\end{aligned} \tag{7.6}
$$

and

$$
\begin{aligned}
\ell_i^c(\boldsymbol{\xi}|y_i, \boldsymbol{x}_i, f_i, \delta_i, w_i) &= \delta_i[\log(\gamma\lambda y_i^{\gamma-1}) + \beta_1 x_{i1} + \beta_2 x_{i2} + f_i] \\
&+ \lambda \exp(\beta_1 x_{i1} + \beta_2 x_{i2} + f_i)(w_i^\gamma - y_i^\gamma) \\
&+ f_i \log\left[\frac{\exp(\alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2})}{1 + \exp(\alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2})}\right] \\
&+ (1 - f_i) \log\left[1 - \frac{\exp(\alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2})}{1 + \exp(\alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2})}\right].
\end{aligned} \tag{7.7}
$$

The score equations to solve are

$$\boldsymbol{u}(\boldsymbol{\xi}) = \sum_{i}^{n} \sum_{f_i} p_{f_i} \times \begin{bmatrix} u_{\gamma,\,i}(\gamma) = \frac{\partial}{\partial \gamma} \ell_i(\boldsymbol{\xi} | y_i, \boldsymbol{x}_i, \delta_i, w_i) \\ u_{\lambda,\,i}(\lambda) = \frac{\partial}{\partial \lambda} \ell_i(\boldsymbol{\xi} | y_i, \boldsymbol{x}_i, \delta_i, w_i) \\ u_{\beta_1,i}(\beta_1) = \frac{\partial}{\partial \beta_1} \ell_i(\boldsymbol{\xi} | y_i, \boldsymbol{x}_i, \delta_i, w_i) \\ u_{\beta_2,i}(\beta_2) = \frac{\partial}{\partial \beta_2} \ell_i(\boldsymbol{\xi} | y_i, \boldsymbol{x}_i, \delta_i, w_i) \\ u_{\alpha_0,i}(\alpha_0) = \frac{\partial}{\partial \alpha_0} \ell_i(\boldsymbol{\xi} | y_i, \boldsymbol{x}_i, \delta_i, w_i) \\ u_{\alpha_1,i}(\alpha_1) = \frac{\partial}{\partial \alpha_1} \ell_i(\boldsymbol{\xi} | y_i, \boldsymbol{x}_i, \delta_i, w_i) \\ u_{\alpha_2,i}(\alpha_2) = \frac{\partial}{\partial \alpha_2} \ell_i(\boldsymbol{\xi} | y_i, \boldsymbol{x}_i, \delta_i, w_i) \end{bmatrix} = \mathbf{0}, \tag{7.8}$$

where p_{f_i} is given by (7.6). The score functions in (7.8) are given by (7.9)-(7.15) at page 77. The EM algorithm takes the following form:

Initialization

1. Estimate $\boldsymbol{\theta}^0 = [\boldsymbol{\beta}^0, \boldsymbol{\eta}^0]$ by maximizing the log-likelihood *without* the frailty component with respect to $\boldsymbol{\theta}$.

$$\boldsymbol{\theta}^0 = \arg\max_{\boldsymbol{\theta}} \left[\sum_i \ell_i(\boldsymbol{\theta} | y_i, \boldsymbol{x}_i, \delta_i, w_i) \right]$$

   where $\ell_i(\boldsymbol{\theta} | y_i, \boldsymbol{x}_i, \delta_i, w_i)$ is given by (7.3).

2. Draw $\alpha_0^0, \alpha_1^0, \alpha_2^0$ from a uniform distribution independently.

3. Set $\boldsymbol{\xi}^0 = [\boldsymbol{\beta}^0, \boldsymbol{\eta}^0, \boldsymbol{\alpha}^0]$.

4. Set $\boldsymbol{\xi}^h = \boldsymbol{\xi}^0$

Updating

1. E-step

   - **for** each subject i
   - **for** each possible value of f_i
   - compute the ith log-likelihood contribution (7.5) given the current value $\boldsymbol{\theta}^h$

$$\sum_{f_i} p(f_i|y_i, \boldsymbol{x}_i, \delta_i, w_i, \boldsymbol{\xi}^h) \times \ell_i^c(\boldsymbol{\xi}|y_i, \boldsymbol{x}_i, f_i, \delta_i, w_i, \boldsymbol{\xi}^h).$$

   - compute the expected complete data log-likelihood

$$Q(\boldsymbol{\xi}|\boldsymbol{\xi}^h) = \sum_i \sum_{f_i} p(f_i|y_i, \boldsymbol{x}_i, \delta_i, w_i, \boldsymbol{\xi}^h) \times \ell_i^c(\boldsymbol{\xi}|y_i, \boldsymbol{x}_i, f_i, \delta_i, w_i, \boldsymbol{\xi}^h)$$

2. Estimate $\boldsymbol{\xi}^{h+1}$ by maximizing $Q(\boldsymbol{\xi}|\boldsymbol{\xi}^h)$ with respect to $\boldsymbol{\xi}$,

$$\boldsymbol{\xi}^{h+1} = \arg\max_{\boldsymbol{\xi}} Q(\boldsymbol{\xi}|\boldsymbol{\xi}^h)$$

   - Update the set of parameters setting $\boldsymbol{\xi}^h = \boldsymbol{\xi}^{h+1}$

The algorithm was stopped when the squared difference between the $(h+1)$th and the hth log-likelihood values was less than 10^{-4}. We consider $\boldsymbol{\theta} = [\gamma, \lambda, \beta_1, \beta_2]'$ the parameters of interest.

The variance of the MLEs $\hat{\boldsymbol{\theta}}$ was given by $\boldsymbol{I}_{WF}^{-1}(\hat{\boldsymbol{\theta}}_{WF})$ and $\boldsymbol{I}_{FD}^{-1}(\hat{\boldsymbol{\theta}}_{FD})$ for WF and FD model respectively, where $\boldsymbol{I}(\hat{\boldsymbol{\theta}})$ is the observed information matrix. The variance of $\hat{\boldsymbol{\theta}}_{EM}$ was computed exploiting the estimator $J_n(\hat{\boldsymbol{\theta}})^{-1}V_n(\hat{\boldsymbol{\theta}})J_n$ defined by (6.10). The vectors of score functions $\boldsymbol{u}_{ij}(\hat{\boldsymbol{\theta}})$ and $\boldsymbol{u}_{ij}(\hat{\boldsymbol{\alpha}})$ in (6.11) were given by the sets of functions (7.9)-(7.12) and (7.13)-(7.15) respectively, considering that f_i could take $m = 2$ values indexed by j in (6.11). The p_{ij}s were given by (7.6) computed for each possible value of f_i and for $\hat{\boldsymbol{\xi}} = [\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\alpha}}]$. Similarly, the ℓ_{ij}^cs in (6.12) were given by (7.7).

$$u_\gamma(\gamma) = \sum_{i=1}^{n} \sum_{f_i} p_{f_i} \times \{\delta_i \left[\gamma^{-1} + \log(y_i)\right] + \exp(\beta_1 x_{1i} + \beta_2 x_{2i} + f_i)\lambda \left[w_i^\gamma \log(w_i^\gamma) - y_i^\gamma \log(y_i^\gamma)\right]\}, \tag{7.9}$$

$$u_\lambda(\lambda) = \sum_{i=1}^{n} \sum_{f_i} p_{f_i} \times [\delta_i \lambda^{-1} + \exp(\beta_1 x_{1i} + \beta_2 x_{2i} + f_i) \left(w_i^\gamma - y_i^\gamma\right)], \tag{7.10}$$

$$u_{\beta_1}(\beta_1) = \sum_{i=1}^{n} \sum_{f_i} p_{f_i} \times [\delta_i x_{1i} + \lambda(w_i^\gamma - y_i^\gamma) \exp(\beta_1 x_{1i} + \beta_2 x_{2i} + f_i) x_{1i}] \tag{7.11}$$

$$u_{\beta_2}(\beta_2) = \sum_{i=1}^{n} \sum_{f_i} p_{f_i} \times [\delta_i x_{2i} + \lambda(w_i^\gamma - y_i^\gamma) \exp(\beta_1 x_{2i} + \beta_2 x_{2i} + f_i) x_{2i}] \tag{7.12}$$

$$u_{\alpha_0}(\alpha_0) = \sum_{i=1}^{n} \sum_{f_i} p_{f_i} \times \{1 - [\exp(\alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i})][1 + \exp(\alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i})]^{-1}\}, \tag{7.13}$$

$$u_{\alpha_1}(\alpha_1) = \sum_{i=1}^{n} \sum_{f_i} p_{f_i} \times x_{1i}\{1 - [\exp(\alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i})][1 + \exp(\alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i})]^{-1}\}, \tag{7.14}$$

$$u_{\alpha_2}(\alpha_2) = \sum_{i=1}^{n} \sum_{f_i} p_{f_i} \times x_{2i}\{1 - [\exp(\alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i})][1 + \exp(\alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i})]^{-1}\}. \tag{7.15}$$

7.1.1 Results

We performed 100 independent simulations according to the set described above and the results are presented for each parameter according to the three different models described at the beginning of this section. The full data model results are the reference with respect to which both the WF model and the EM procedure performances are compared. For each parameter's estimate we computed the standard error (SE) and the coverage probability (CP). The SEs were given by the element i_{ss} of the estimated variance covariance matrix, $s = 1, ..., k$ where $k = 4$ is the number of the parameters of interest. The CPs are empirical probabilities given by the frequencies of inclusion of the true value of θ_s in the associated confidence interval (CI). In the current setting the CI associated to the generic parameter $\boldsymbol{\theta}_s$ was given by given by

$$\hat{\theta}_s - z_{\frac{\alpha}{2}}\sqrt{\widehat{Var}(\hat{\theta}_s)} \leq \theta_s \leq \hat{\theta}_s + z_{\frac{\alpha}{2}}\sqrt{\widehat{Var}(\hat{\theta}_s)},$$

where $\widehat{Var}(\hat{\theta}_s)$ is the appropriate variance estimator under the considered model and $z_{\frac{\alpha}{2}}$ is the $\alpha = 0.05$ quantile of Z(0,1). Such a distribution was chosen because of the properties of MLEs for large samples.

Table (7.1) provides the results for the baseline parameters, while table (7.2) provides the results for β_1 and β_2. The γ estimate of WF is more biased than the EM one and has a really low coverage probability (4%) while in the EM case it is 100%, even higher than the full data case (FD) one. The WF performance gets even worse looking at the λ estimates for which the bias is really severe and the coverage probabilities are null. So far, for the the baseline parameters it is evident that EM estimates provided a very good performance, as the results are really close to the full data model, except for a bit larger standard errors. This similarity is also evident looking at boxplots in Figure 7.1.

The results are slightly different for β_1 and β_2. The standard errors are larger in all the tested procedures. As expected the full data analysis provided the

best result as the estimate is correct, the standard error is small and the coverage probability is 92%. The WF case provides a better result for β_1 than the EM procedure: its estimate is less biased and also the standard error is smaller, while the coverage probability is higher (Tab.7.2). The β_2 estimates show a similar output pattern, as the FD performance is as expected the best and the WF results are quite better than the EM ones (Figure 7.2). So far, at a first glance the EM procedure performs better than the WF one for the baseline parameters estimates (γ and λ), while it is worst for the risk parameters (β_1 and β_2). However, it is evident that when it's the WF procedure to perform worse respect to the EM, namely for γ and λ, the bias is really higher and the coverage probabilities are really lower with respect to the case when it is the EM to perform worse (β_1 and β_2). Hence, we can state that on the whole the EM procedure provides better results than the complete case one. As a further result the EM correctly imputed 64% of missing values.

Table 7.1: Simulation results: baseline parameters.

	Estimate (Standard Error)		Coverage (95%)	
Case	$\gamma = 0.6$	$\lambda = 0.3$	γ	λ
FD	0.60 (.02)	0.30 (.01)	.93	.90
WF	0.54 (.02)	0.53 (.02)	.04	.00
EM	0.58 (.04)	0.29 (.03)	1.00	.91

Table 7.2: Simulation results: risk parameters.

Estimate (Standard Error)			Coverage (95%)	
Case	$\beta_1 = 0.7$	$\beta_2 = 0.5$	β_1	β_2
FD	0.70 (.04)	0.49 (.04)	.92	.85
WF	0.68 (.04)	0.56 (.04)	.89	.56
EM	0.75 (.06)	0.60 (.06)	.76	.45

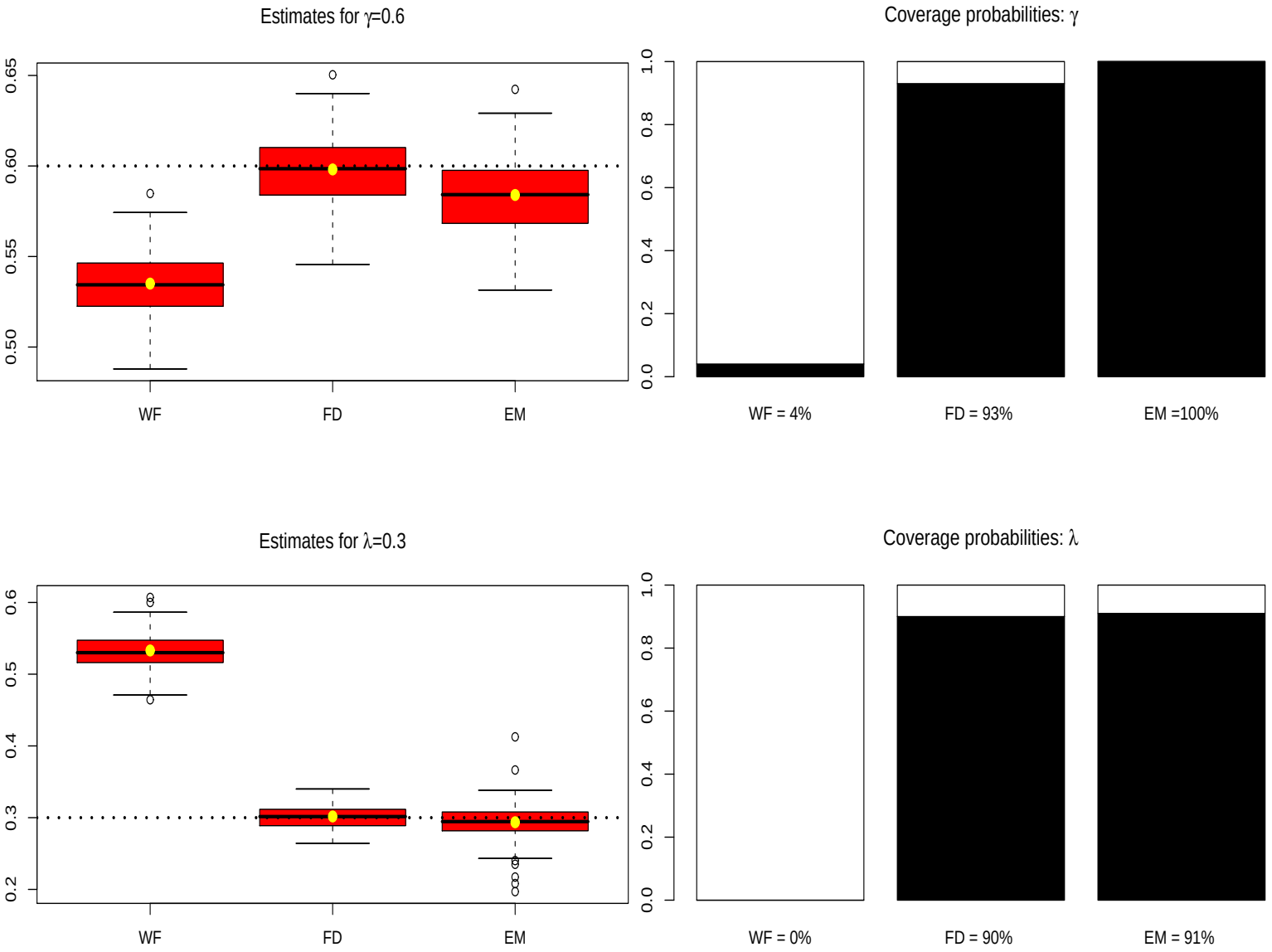


Figure 7.1: Box plots and coverage probabilities for γ and λ.

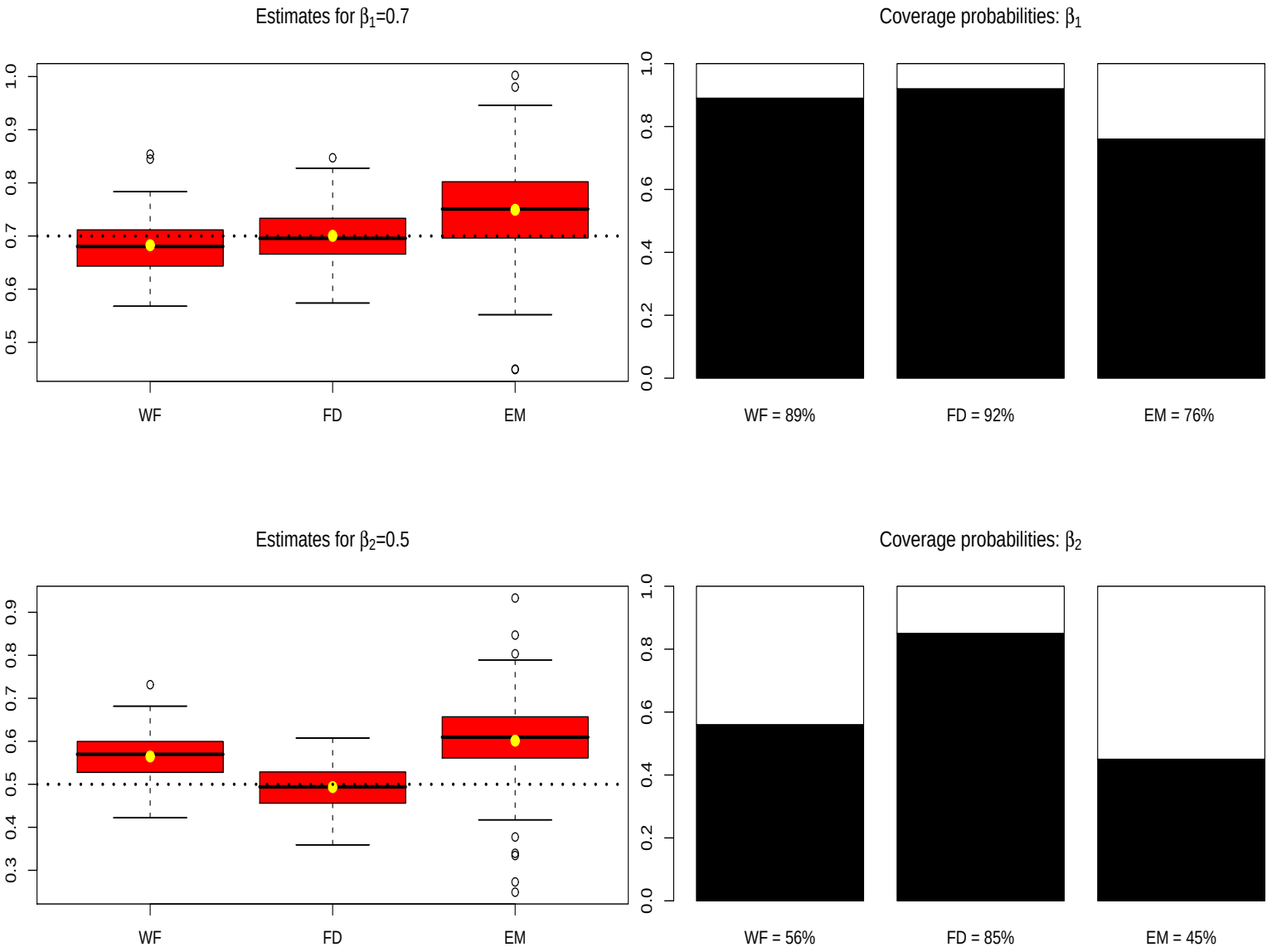


Figure 7.2: Box plots and coverage probabilities for β_1 and β_2.

7.2 Continuous random effect

We built up a set of $n = 500$ subjects each described by one covariate x_i which is observed $\forall i$ and was independently drawn from a random variable $X \sim N(0, 1)$. We assumed an exponential baseline hazard,

$$h_0(t|\eta) = \lambda \tag{7.16}$$

where $\lambda = 1$. Each subject was assigned a random effect f_i. Each random effect f_i was independently drawn from a subject-specific distribution. In this setting it was

$$f_i \sim N(0, \sigma_i^2)$$

where we assumed

$$\sigma_i^2 = \exp[\alpha_0 + \alpha_1(x_i - \bar{x})]$$

and $\alpha_0 = 0.2$, $\alpha_1 = 0.3$. We generated n event times $\{t_i\}$ randomly, according to the subject-specific conditional hazard

$$h(t_i|x_i, f_i, \boldsymbol{\theta}) = \lambda \exp(\beta_1 x_i + f_i), \qquad \boldsymbol{\theta} = [\lambda, \beta]'$$

setting $\beta = 0.7$. Namely it was

$$t_i = -\log(1 - u)[\lambda \exp(\beta x_i + f_i)]^{-1}$$

where $u \sim U(0, 1)$ is the random value of the cumulative distribution function of the lifetime t_i given by

$$F(t_i|\boldsymbol{x}_i, f_i, \boldsymbol{\theta}) = 1 - S(t_i|\boldsymbol{x}_i, f_i, \boldsymbol{\theta}) = 1 - \exp[-\lambda t_i \exp(\beta x_i + f_i)].$$

If we do not consider the presence of the random effect f_i we define the ith log-likelihood contribution as

$$\ell_i(\boldsymbol{\theta}|t_i, x_i) = \log\lambda + \beta x_i - \lambda t_i \exp(\beta x_i) \tag{7.17}$$

Assuming we know the real value of $f_i\,\forall i$ the ith log-likelihood contribution is

$$\ell_i(\boldsymbol{\theta}|t_i, x_i, f_i) = \log\lambda + \beta x_i + f_i - \lambda t_i \exp(\beta x_i + f_i)$$

under the full data model (FD).

Recalling that we do not know the value of f_i actually we need to consider the prior density of f_i

$$p(f_i|x_i, \boldsymbol{\alpha}) = \frac{1}{\sqrt{2\pi\exp(\alpha_0 + \alpha_1 x_i)}} \exp\left[-\frac{f_i^2}{2\exp(\alpha_0 + \alpha_1 x_i)}\right]$$

to specify the model correctly. We may write the ith log-likelihood contribution as an expected complete data log-likelihood contribution,

$$E[\ell_i(\boldsymbol{\xi}|t_i, x_i, f_i)|t_i, x_i] =$$

$$\int_{f_i} p(f_i|t_i, x_i, \boldsymbol{\xi}) \times \ell_i^c(\boldsymbol{\xi}|t_i, x_i, f_i) df_i, \tag{7.18}$$

where

$$p(f_i|t_i, x_i, \boldsymbol{\xi}) =$$

$$\frac{\lambda\exp(\beta x_i + f_i)\exp[-_i\exp(\beta x_i + f_i)][2\pi e^{\alpha_0+\alpha_1(x_i-\overline{x})}]^{-\frac{1}{2}} e^{-f_i^2\{2\cdot\exp[\alpha_0+\alpha_1(x_i-\overline{x})]\}^{-1}}}{\int_{f_i}\lambda\exp(\beta x_i + f_i)\exp[-\lambda t_i\exp(\beta x_i + f_i)][2\pi e^{\alpha_0+\alpha_1(x_i-\overline{x})}]^{-\frac{1}{2}} e^{-f_i^2\{2\cdot\exp[\alpha_0+\alpha_1(x_i-\overline{x})]\}^{-1}} df_i}. \tag{7.19}$$

and

$$\begin{aligned}\ell_i^c(\boldsymbol{\xi}|t_i, x_i, f_i) &= \log\lambda + \beta x_i + f_i - \lambda t_i \exp(\beta x_i + f_i) - \frac{1}{2}[\log 2\pi + \alpha_0 + \alpha_1(x_i - \overline{x})] \\ &- \frac{1}{2} f_i^2 \{\exp[\alpha_0 + \alpha_1(x_i - \overline{x})]\}^{-1} \end{aligned} \tag{7.20}$$

In the current setting we assumed to know the value of λ, so the only parameter we need to estimate is β. Because of this, under the WF and FD models assumptions we need to solve the score equation

$$u(\beta) = \sum_i^n \ell_i'(\beta|y_i, x_i) = 0 \tag{7.21}$$

only. Under the WF and the FD models the ith contribution to the score function in (7.21) are given by

$$u_\beta(\beta) = x_i \cdot [1 - \lambda \exp(\beta x_i)],$$

and

$$u_\beta(\beta) = x_i \cdot [1 - \lambda \exp(\beta x_i + f_i)]$$

respectively.

We exploited the MCEM algorithm as described in Section 6.3.2in order to estimate β accounting for the unobserved random effect f_i. This implies that the expected complete data log-likeliood contribution (7.18) was approximated by

$$\widetilde{E}[\ell_i^c(\boldsymbol{\xi}|t_i, x_i, f_i)|t_i, x_i] = \frac{1}{m} \sum_{r=q+1}^{R} \ell_i^c(\boldsymbol{\xi}|t_i, x_i, s_{i,r}). \tag{7.22}$$

where $\ell_i^c(\boldsymbol{\xi}|t_i, x_i, s_{i,r})$ is given by (7.19) replacing f_i with $s_{i,r}$. The quantity $s_{i,r}$ is the rth component of the p-vector of sampled values of f_i drawn from (7.19) via the adaptive rejection sampling algorithm (Gilks and Wild, 1992)

described in Section 4.3.1. Note that the posterior density (7.19) from which we sample $\{s_i\}$ is univariate, so the whole sampling procedure can be seen as a particular case of Gibbs sampler where we have only one full conditional density to sample from (see Section 4.3). After checking for the convergence of Gibbs sampler we chose a burn-in period of $q = 1000$ and a sample size of $m = p - q = 200$. According to (7.22) the score equations to solve are therefore

$$\boldsymbol{u}^*(\boldsymbol{\xi}) = \frac{1}{m}\sum_{i}^{n}\sum_{r=q+1}^{R}\begin{bmatrix} u^*_{\beta,\,i}(\beta) = \frac{\partial}{\partial\beta}\ell_i^c(\boldsymbol{\xi}|t_i, x_i, s_{i,r}) \\ u^*_{\alpha_{0,\,i}}(\alpha_0) = \frac{\partial}{\partial\alpha_0}\ell_i^c(\boldsymbol{\xi}|t_i, x_i, s_{i,r}) \\ u^*_{\alpha_{1,\,i}}(\alpha_1) = \frac{\partial}{\partial\alpha_1}\ell_i^c(\boldsymbol{\xi}|t_i, x_i, s_{i,r}) \end{bmatrix} = \mathbf{0},$$

where the ith contributions to the score functions are given by

$$\begin{aligned} u^*_\beta(\beta) &= x_i \cdot [1 - t_i \exp(\beta_1 x_i + s_{i,r})], \\ u^*_{\alpha_0}(\alpha_0) &= 0.5 \cdot \{(s^h_{i,r})^2 \cdot \exp[-\alpha_0 - \alpha_1(x_i - \overline{x})] - 1\}, \\ u^*_{\alpha_1}(\alpha_1) &= 0.5 \cdot (x_i - \overline{x}) \cdot \{(s^h_{i,r})^2 \cdot \exp[-\alpha_0 - \alpha_1(x_i - \overline{x})] - 1\}. \end{aligned}$$

The Monte Carlo EM algorithm takes the following form:

Initialization

1. Estimate β^0 maximizing the log-likelihood *without* the frailty component with respect to β.

$$\beta^0 = \arg\max_{\beta}\left[\sum_i \ell_i(\beta|t_i, x_i)\right]$$

where

$$\ell_i(\beta|t_i, x_i) \quad = \quad \beta x_i - t_i \exp \beta x_i$$

2. Draw α_0, α_1 from a uniform distribution.

3. Set $\boldsymbol{\xi}^0 = [\beta^0, \alpha_0^0, \alpha_1^0]$

4. Set $\boldsymbol{\xi}^h = \boldsymbol{\xi}^0$

5. Set $q = 1000$ as the burn-in period for the Gibbs sampler.

Updating

1. E-step:

   - **for** each subject i

   - Sample $m = p - q$ values for f_i from $p(f_i|t_i, x_i, \boldsymbol{\xi}^h)$ given by (7.19) setting $\lambda = 1$ and $\boldsymbol{\xi}^h$ is the current estimate of $\boldsymbol{\xi}$.

   - compute the approximated expected complete data log-likelihood conribution

$$\widetilde{E}[\ell_i^c(\boldsymbol{\xi}|t_i, x_i, f_i)|t_i, x_i] = \frac{1}{m} \sum_{r=q+1}^{R} \ell_i^c(\boldsymbol{\xi}|t_i, x_i, s_{i,r}).$$

   given $\boldsymbol{\xi}^h$.

- compute the expected complete data log-likelihood

$$Q(\boldsymbol{\xi}|\boldsymbol{\xi}^h) = \frac{1}{m}\sum_i \sum_{r=q+1}^{R} \ell_i^c(\boldsymbol{\xi}|t_i, x_i, s_{i,r})$$

2. M-step: estimate $\boldsymbol{\xi}^{h+1}$ by maximizing $Q(\boldsymbol{\xi}|\boldsymbol{\xi}^h)$ with respect to $\boldsymbol{\xi}$

$$\boldsymbol{\xi}^{h+1} = \arg\max_{\boldsymbol{\xi}} Q(\boldsymbol{\xi}|\boldsymbol{\xi}^h)$$

- Update the set of parameters setting $\boldsymbol{\xi}^h = \boldsymbol{\xi}^{h+1}$.

The convergence criterion for the MCEM algorithm was that the squared difference between the $(h+6)$th and the hth log-likelihood values was less than 10^{-3}.
The variance of the MLEs $\hat{\beta}$ were given by $\boldsymbol{i}_{WF}^{-1}(\hat{\beta}_{WF})$ and $\boldsymbol{i}_{FD}^{-1}(\hat{\beta}_{FD})$ for WF and FD model respectively, where $\boldsymbol{i}^{-1}(\beta)$ is the observed information number. The variance of $\hat{\boldsymbol{\theta}}_{EM}$ was computed exploiting the estimator $\hat{V} = \bar{W} + (1+m^{-1})B$ given by (6.16). Given the MLE estimate $\hat{\beta}$ we drew $m = 20$ values for each f_i from $p(f_i|x_i, f_i, \hat{\beta}) =$ where $\hat{\beta}$ is the MLE estimate after the MCEM was stopped. We built up an $n \times m$ matrix $\boldsymbol{S}$ where the jth column is the set of n sampled values for the random effect f. Then we estimated m different vectors of parameters $\hat{\boldsymbol{\theta}}_j$ by maximizing m different log-likelihoods

$$\ell_j(\beta|\boldsymbol{t}, \boldsymbol{x}, \boldsymbol{S}) \;=\; \sum_{i=1} \ell^c(\beta|t_i, x_i, s_{ij}) \tag{7.23}$$

where $j = 1, 2, ..., m$ indexes the columns of $\boldsymbol{S}$. The ith log-likelihood contribution in (7.23) is given by

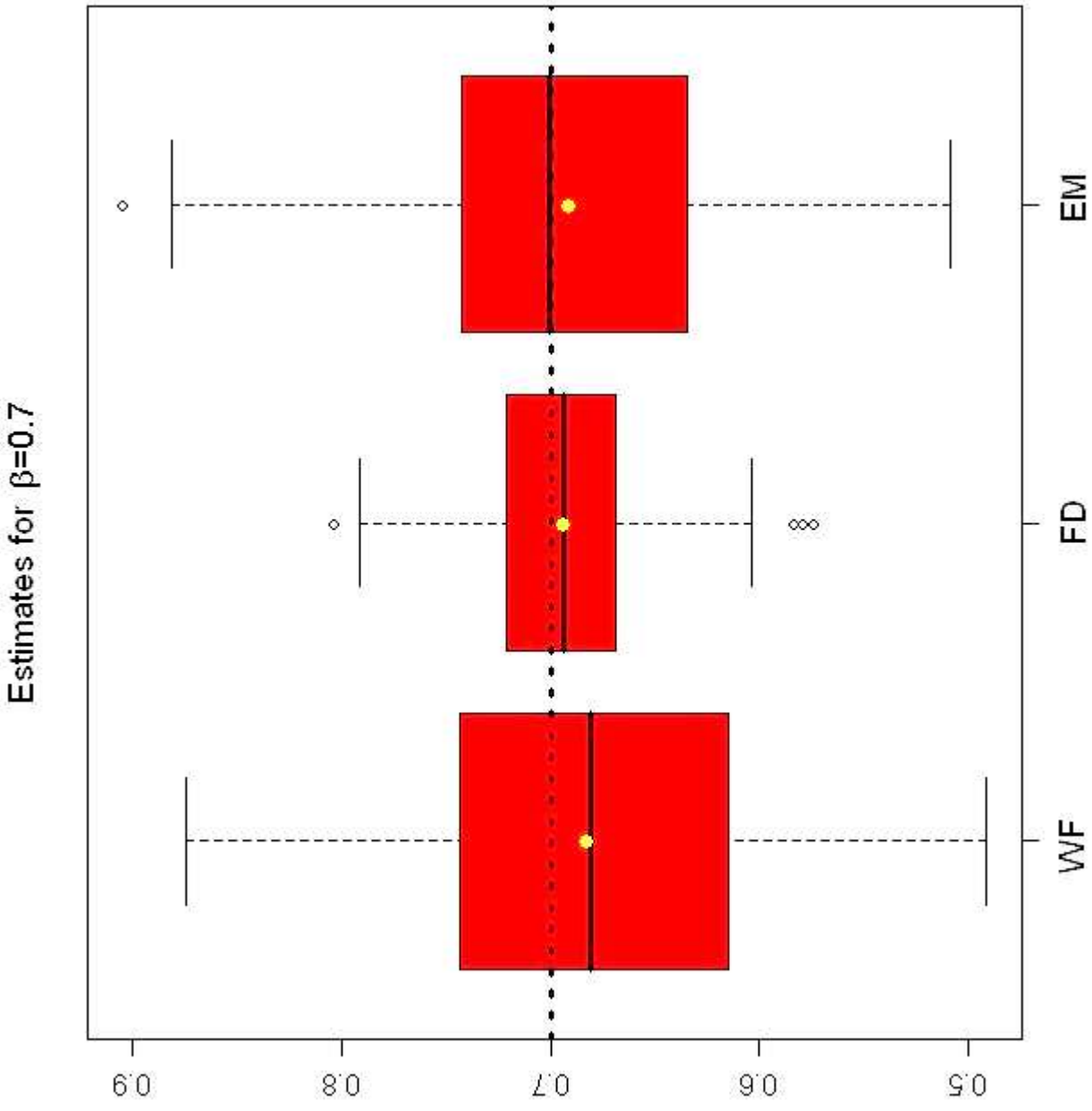
$$\sum_{i=1} \ell^c(\beta|t_i, x_i, s_{ij}) \sum_{i=1}^{n} \beta x_i + s_{ij} - t_i \exp(\beta x_i + s_{ij})$$

where s_{ij} is the jth imputed value of f_i. The final result is a set of $m = 20$ MLE estimates $\hat{\beta}_j^{IMP}$ which we use to compute the estimator (6.16)

The MCEM procedure provides a good estimate of β as shown in Tab.7.3 and Figure 7.3. The estimates given by MCEM are the same of the full data model, providing even higher coverage probabilities.

Table 7.3: Simulation results: risk parameter.

Estimate (Standard Error)		Coverage (95%)
Model	$\beta = 0.7$	C.P.
FD	0.69 (.04)	.91
WF	0.68 (.04)	.55
EM	0.69 (.06)	.94


Estimates for β=0.7
0.5
0.6
0.7
0.8
0.9
EM
FD
WF

Chapter 8

Conclusions and outlook

Survival analysis focuses on the duration between a time origin and the occurrence of an event of interest. The main analysis' tools are the survival function and the hazard function. For each observed subject the survival function provides the probability that the event occurs after a given time point t. The hazard function provides the risk that the event occurs an instant after time t given it has not occurred yet. The time range of our observation may not coincide with the duration between time origin and event. Thus, different observation schemes are possible. One of these includes both possible delayed entries in the observation and an exits from the observation before the event occurs. We name such occurrences as left-truncation and right-censoring respectively. The presence of possible left-truncation and right-censoring affects the specification of the model in terms of the hazard function, the survival function and of the likelihood function as well. Each subject has peculiar features which impact on his susceptibility to the event's occurrence which make each individual different. In other words there is heterogeneity among the individuals belonging to a given population. Part of such heterogeneity may be described by fixed effects models (FE), i.e. by describing each subject by a function of a given set of covariates named *predictor*. The heterogeneity which is captured by fixed effects is named *observed* heterogeneity. A very popular FE model is the proportional hazards model

(PH), in which the individual hazard is made up of two components. One is a shared component named baseline, which is multiplied by the exponential of a linear combination of *observable* covariates which are assumed to be fixed.

It could be possible that unobservable subject or group-specific features impact on the individual hazard functions. This implies that in a given population could be several sub-groups with different hazard functions. Such phenomenon generates the *unobserved* heterogeneity. If we do not account for it the hazard of the population would be underestimated while the population's survival function would be overestimated (Omori and Johnson, 1993). Such *unobserved* heterogeneity is usually modeled by *random effects*, which are the unobservable realizations of a random variable. In survival analysis the random effect is named *frailty*, the term which Vaupel et al. (1979) used to denote a subject-specific fixed component which was unobservable and multiplied by the baseline hazard rate. A model which accounts for both observed and unobserved heterogeneity is the proportional hazards frailty model i which the individual hazard is made up of three components (Clayton and Cuzick, 1985). The first one is the baseline. The second component is the subject-specific predictor previously defined. The third component is the frailty. This is our refernce model

Although the original frailty model has been widely extended and generalized they have been defined as a quantity which is independent from fixed effects The key idea behind this work is that an interaction between random effects and fixed effects may exist. Such interaction has been modeled assuming that the distribution of each subject-specific random effects has the same functional form the parameters of which depends on the values of one or more observed covariates. This assumption is general with respect to the different kinds of survival models. In this work we included it in an over-dispersion proportional hazards model assuming a discrete and a continuous random effect alternatively. The random effect was dealth with as a missing covariate for each subject, thus we proposed the estimation procedures presented by Herring and Ibrahim (2001) for

a proportional hazard model with missing covariates. We performed two different sets of simulations in which these models were estimated by the EM and MCEM algorithm respectively. In each set we compared the results of such procedures with those of two other models, one wich did not account for unobserved heterogeneity (WF), the other where the value od random effect was assumed to be known (FD). For both sets of simulations the EM and MCEM algorithms provided estimates near to those of the full data model.

Although the assumption of an interaction between fixed and random effects is rather general, the model we tested in this work is only a part of the possible survival models which could be implemented relying on it. For example an immediate generalization could be made by including our assumption in a semi-parametric proportional hazards frailty model. A more challenging generalization could be the introduction of missing values in the covariates, or (and) allowing for time-varying covariates.

Bibliography

Agarwal, R. and D. B. Audretsch (2001, March). Does entry size matter? the impact of the life cycle and technology on firm survival. *The Journal of Industrial Economics 49*(1), 21–43.

Agarwal, R. and M. Gort (2002, May). Firm and product life cycles and firm survival. *The American Economic Review 92*(2), 184–190.

Belzil, C. (2001, September). Unemployment insurance and subsequent job duration: Job matching versus unobserved heterogeneity. *Journal of Applied Econometrics 16*(5), 619–636.

Bennett, P., R. Peach, and S. Peristiani (2001, November). Structural change in the mortgage market and the propensity to refinance. *Journal of Money, Credit and Banking 33*(4), 955–975.

Clayton, D. and J. Cuzick (1985). Multivariate generalizations of the proportional hazards model. *Journal of the Royal Statistical Society, Series A 148*(2), 82–117.

Cockx, B. and G. Ridder (2001, April). Social employment of welfare recipients in belgium: An evaluation. *The Economic Journal 111*(470), 322–352.

Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society, Series B 34*(2), 187–202.

Cox, D. R. and D. Oakes (1984). *Analysis of Survival Data*. Chapman & Hall/CRC.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B 39*, 1–38.

Fisher, R. A. (1925). Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society 22*, 700–725.

Gamerman, D. and M. West (1987). An application of dynamic survival models in unemployment studies. *The Statistician 36*(2/3), 269–274. Special Issue: Practical Bayesian Statistics.

Gelfand, A. E. and A. F. Smith (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association 85*, 398–409.

Geman, S. and D. Geman (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence 6*, 721–741.

Gilks, W. R., S. Richardson, and D. J. Spiegelhalter (1998). *Markov chain Monte Carlo in practice* (1st ed.). Chapman-Hall/CRC. First CRC Press reprint 1998.

Gilks, W. R. and P. Wild (1992). Adaptive rejection sampling for gibbs sampling. *Applied Statistics 41*(2), 337–348.

Goetgheburger, E. and L. Ryan (2000, December). Semiparametric regression analysis of interval-censored data. *Biometrics 56*, 1139–1144.

Guiso, L. and T. Jappelli (2002, May). Private transfers, borrowing constraints and the timing of homeownership. *Journal of Money, Credit and Banking 34*(2), 315–339.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika 57*, 97–109.

Herring, A. H. and J. G. Ibrahim (2001). Likelihood-based methods for missing covariates in the Cox proportionalhazards model. *Journal of American Statistic Association 85*, 292–302.

Herring, A. H., J. G. Ibrahim, and S. R. Lipsitz (2002, March). Frailty models with missing covariates. *Biometrics 58*, 98–109.

Hougaard, P. (1984). Life table methods for heterogeneous populations: Distributions describing the heterogeneity. *Biometrika 71*(1), 75–83.

Hougaard, P. (1986, August). Survival models for heterogeneous populations derived from stable distributions. *Biometrika 73*(2), 387–396.

Hougaard, P. (2001). *Analysis of Multivariate Survival Data* (2nd ed.). Springer.

Ibrahim, J. G., M. H. Chen, and S. N. MacEachern (1999, December). Bayesian variable selection for proportional hazards models. *The Canadian Journal of Statistics 27*(4), 701–707.

Kiefer, N. M. (1988, June). Economic duration data and hazard functions. *Journal of Economic Literature 26*(2), 646–679.

Klein, J. P. and M. L. Moeschberger (1997). *Survival analysis* (1 ed.). Springer-Verlag, New York.

Lancaster, T. (1979, July). Econometric methods for the duration of unemployment. *Econometrica 47*(4), 939–956.

Lancaster, T. (1992, September). *The Econometric Analysis of Transition Data*. Number 17 in Econometric Society Monographs. Cambridge University Press.

Li, H. and Y. Xu (2002, October). Survival bias and the equity premium puzzle. *The Journal of Finance 57*(5), 1981–1995.

Lipsitz, S. R. and J. G. Ibrahim (1996a). A conditional model for incomplete covariates in parametric regression models. *Biometrika 83*(4), 916–922.

Lipsitz, S. R. and J. G. Ibrahim (1996b). Using the em-algorithm for survival data with incomplete categorical covariates. *Lifetime Data Analysis 2*, 5–14.

Little, J. A. and D. B. Rubin (1987). *Statistical analysis with missing data*.

Mata, J. and P. Portugal (1994, September). Life duration of new firms. *The Journal of Industrial Economics 42*(3), 227–245.

McLachlan, G. and T. Krishnan (1996, November). *The EM Algorithm and its Extensions* (1 ed.). Wiley Series in Probability and Statistics. Wiley-Interscience.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller (1953). Equations of state calculations by fast computing machine. *Journal of Chemical Physics 21*, 1087–1091.

Omori, Y. and R. A. Johnson (1993). The influence of random effects on the unconditional hazard rate and survival functions. *Biometrika 80*(4), 910–914.

Petersen, J. H. (1998, June). An additive frailty model for correlated life times. *Biometrics 54*(2), 646–661.

Prabhakar Murthy, D. N., M. Xie, and R. Jiang (2003). *Weibull models*. Wiley Series in Probability and Statistics. John Wiley & Sons Inc.

Pugh, M., J. Robins, S. Lipsitz, and D. Harrington (1993). Inference in the Cox Proportional Hazards Model with MissingCovariate Data. Technical report, Dana-Farber Cancer Institute, Boston, Division of Biostatistical Science.

Rosin, P. and B. Rammler (1933). The laws governing the fineness of powdered coal. *Journal of the Institute of Fuel 6*, 29–36.

Royston, G. H. D. (1983, September). Wider application of survival analysis: An evaluation of an unemployment benefit. *The Statistician 32*(3), 302–306.

Rubin, D. and N. Schenker (1991). Multiple imputation in health-care data bases: an overview and some applications. *Statistics in medicine 10*, 585–598.

Rubin, D. B. (1976). Inference and missing data. *Biometrika 63*(3), 581–592.

Schafer, J. L. and J. W. Graham (2002). Missing data: our view of the state of the art. *Psychological methods 7*(2), 147–167.

Vaida, F. and R. Xu (2000). Proportional hazards model with random effects. *Statist. Med. 19*, 3309–3324.

Vaupel, J. W., K. G. Manton, and E. Stallard (1979, August). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography 16*(3), 439–454.

Weibull, W. (1939). A statistical theory of the strength of materials. *Ingeniors Vetenskaps Akademien Handlingar (Proceedings of the Royal Swedish Institute for Engineering Research) 151*, 1–45.

Weibull, W. (1951). A statistical distribution of wide applicability. *Journal of Applied Mechanics 18*, 293–297.

Yue, H. and K. S. Chan (1997, September). A dynamic frailty model for multivariate survival data. *Biometrics 53*(3), 785–793.